A Formal Account of Dishonesty

Chiaki Sakama
Department of Computer and Communication Sciences
Wakayama University, Sakaedani, Wakayama 640-8510, Japan
Email: sakama@sys.wakayama-u.ac.jp

Martin Caminada
Department of Computing Science
University of Aberdeen, Aberdeen AB24 3UE, United Kingdom
Email: martin.caminada@abdn.ac.uk

Andreas Herzig
University of Toulouse and CNRS, France
Email: herzig@irit.fr

November 2014

Abstract

This paper provides formal accounts of dishonest attitudes of agents. We introduce a propositional multi-modal logic that can represent an agent's *belief* and *intention* as well as *communication* between agents. Using the language, we formulate different categories of dishonesty. We first provide two different definitions of *lies* and provide their logical properties. We then consider an incentive behind the act of lying and introduce lying with objectives. We subsequently define *bullshit*, *withholding information* and *half-truths*, and analyze their formal properties. We compare different categories of dishonesty in a systematic manner, and examine their connection to *deception*. We also propose *maxims* for dishonest communication that agents should ideally try to satisfy.

Keywords: dishonesty, multi-modal logic, belief and intention, communication

1 Introduction

> *Men are able to trust one another, knowing the exact degree of dishonesty they are entitled to expect.*
>
> —Stephen Leacock (1869–1944, Canadian Humorist)

Dishonesty is a fact of daily life. A behavioral economist Dan Ariely experimentally shows that there are various forces which urge people to behave dishonestly [3]. He argues that an average person has "the delicate balance between the contradictory desires to maintain a positive self-image and to benefit from cheating", and a person rationalizes his/her dishonest behavior using this balancing act. There are a number of words representing dishonest attitudes of humans—bluff, bullshit, cheat, concealing, deceit, dodge, fake, fraud, fudge, and lie, to name a few. These different categories of dishonesty frequently appear in human communication to benefit the self or avoid conflict. According to a study in social psychology [26], 20%–30% of conversation are in fact lies.

In classical *speech act theory*, on the other hand, a speaker who makes an assertion is assumed to obey *the sincerity rule* that "the speaker commits himself to a belief in the truth of the expressed proposition" [63, p.62]. This basic principle is not applied to human agents who may behave dishonestly in communication. The issue is also pointed out for the FIPA agent communication language: "In Arcol, an agent must make only *sincere contributions* (assertives that are believed true, requests that it intends should succeed) and may assume that other agents also make only sincere contributions. Consequently, you cannot use Arcol in settings where sincerity cannot be taken for granted—for example, in electronic commerce or, broadly, in negotiation of any kind" [65]. The need for relaxing the sincerity condition in agent communication is also argued for in [51]. As such, formulating dishonesty in agent communication and constructing behavioral rules for dishonest agents are necessary and challenging. Many philosophers have discussed dishonest behaviors of humans in the literature [7, 15, 33, 39, 64, 67, 71], while relatively little study exists for formulating and comparing different categories of dishonesty. A reason for this is the fact that defining the notion of dishonesty has been the topic of extensive discussion. For instance, Mahon [42] examines twelve different definitions of lying in the literature and argues which one is empirically acceptable. Moreover, when we reason about dishonesty we cannot simply take a subjective view and represent an agent's beliefs by means of formulas of classical propositional logic. We need modal operators of belief and intention in order to be able to represent at the same time what an agent believes and what the agent intends to communicate (which might contradict the agent's beliefs). Thus, providing a formal account of dishonesty requires one to overcome various difficulties.

The issue of formulating dishonesty is relevant as a research topic not only in philosophy, but also in *artificial intelligence* (AI). In AI the question "Can computers deceive humans?" has been of interest since Turing's *imitation game* [72]. The purpose of the game is to devise an *intelligent* computer which behaves like humans in conversation. A computer attempts to convince a judge that it is human through appropriate, and often *deceptive*, responses. A recent study shows that lying by computers

affects judges' misclassification in that computers are considered human [80].[1] There are several reasons why the study of dishonesty is important in AI. (i) First, dishonest behaviors are inherent to human beings, that require intelligence and thinking.[2] Some researchers remark this fact: "Lying is related to intelligence ... lying demands both advanced cognitive development and social skills that honesty simply doesn't require" [10] and "lying is challenging in terms of cognitive control because an individual must hold two mental states (i.e., their own and that of another) in mind" [13]. Studies on dishonesty can therefore contribute to better understanding human intelligence and take us one step closer to realizing "human-like" AI. (ii) Second, understanding the mechanism of dishonesty opens possibilities to develop computers which select dishonest behaviors as moral or secretive decisions. Such "artificial liars" are considered effective in a number of situations [16]. For instance, we can imagine a robot for medical care which does not inform a patient of the true state of affairs. An intelligent personal assistant might deceive us to influence us to make a right decision. Applications of lying in AI and knowledge engineering are also reported in the literature [8, 17, 27, 66, 78]. (iii) Third, studying dishonest acts in the context of multiagent systems is necessary for designing social agents who behave economically to minimize costs and/or maximize benefits. Lying or deception is in fact one of the strategic interactions between self-interested agents, and its effects have been analyzed from the game-theoretic viewpoints [24, 28, 75]. Some studies show potential utilities of dishonest acts as a strategy in multiagent negotiation [28, 69, 81], formal argumentation [11, 53, 59], and in social interactions [12, 70]. (iv) Fourth, the new field of *computational morality* or *machine ethics* has emerged in AI for computationally modelling moral decision making [2, 79]. Formal theories for encoding ethical rules and computational logics for realizing computational morality have also been studied by several researchers [34, 50]. Clearly, morality and dishonesty are closely related—they are two different sides of the same coin. Studying dishonesty will thus contribute to better understanding morality and designing ethical agents.

The purpose of this paper is to provide a logical account of dishonesty. To represent dishonest attitudes of agents, we need a logic that can distinguish belief of truth and an act of falsehood. Moreover, it is necessary to represent intention of a speaker and to reason about the belief state of a hearer. To this end, we first introduce a propositional multi-modal logic **BIC** that can represent three modalities: *belief*, *intention* and *communication*. Using the logic, we first formulate sincere communication between agents, which is later contrasted with dishonest communication. We then proceed to formal accounts of dishonesty where four different categories are considered: *lie*, *bullshit*, *withholding information* and *half-truth*. We represent them using the logical language of **BIC** and prove their semantic properties using its axiomatic system. We compare those categories in a systematic manner and argue their connection to *deception*. We also address qualitative and quantitative *dishonesty maxims* that agents should try to satisfy both for moral and self-interested reasons.

The rest of this paper is organized as follows. Section 2 introduces the logic that is

[1] Interestingly, an experiment shows that truth-telling by a machine is often taken as a double bluff by a human judge resulting in misclassification it as human.

[2] It is well known that non-human animals often deceive other individuals, which is a result of natural selection for the struggle for existence [45].

used in this paper and formulates sincere communication. Section 3 provides a logical framework of lies and investigates formal properties. Section 4 formulates bullshit, withholding information, and half-truth. Section 5 proposes postulates and maxims for dishonest agents. Section 6 compares different categories of dishonesties and addresses related works. Section 7 concludes the paper.

2 A Logic for Belief, Intention and Communication

In this section, we first introduce a propositional multi-modal logic **BIC** which is used in this paper. We then formulate in **BIC** conditions for sincere communication.

2.1 BIC logic

The propositional multi-modal language $\mathcal{L}$ of the logic **BIC** is built from a finite set of propositional constants $P = \{p, q, r, \ldots\}$ on the logical connectives $\neg$ and $\wedge$, and on three different types of modal operators, $(B_a)_{a\in A}$, $(I_a)_{a\in A}$, and $(C_{ab})_{a,b\in A}$, where $A = \{a, b, c, \ldots\}$ is a finite set of *agents*.[3] Well-formed formulas (or *sentences*) in $\mathcal{L}$ are defined as follows: (i) If $p \in P$, then p is a sentence. (ii) If φ and ψ are sentences, then $\neg\varphi$ and $\varphi \wedge \psi$ are sentences. (iii) If φ is a sentence and $a, b \in A$, then $B_a\varphi$, $I_a\varphi$, and $C_{ab}\varphi$ are all sentences. The logical connectives $\top$, $\bot$, $\vee$, $\supset$ and $\equiv$ are introduced as abbreviations as usual. The set of all sentences in $\mathcal{L}$ is denoted by Φ. Throughout the paper, lower case letters a and b represent agents in A and Greek letters $\delta, \lambda, \sigma, \varphi, \psi$ represent sentences in Φ unless otherwise stated. A sentence $B_a\varphi$ represents an agent's internal belief and is read as "an agent a believes φ". A sentence $I_a\varphi$ represents an agent's intention in action and is read as "a intends φ". A sentence $C_{ab}\varphi$ represents an agent's intentional communication to another agent and is read as "a communicates φ to b". By communication, we understand any method by which an agent a informs an agent b of a sentence σ, be it linguistically or not. Nevertheless, we often call "a" a *speaker* and "b" a *hearer* in $C_{ab}\sigma$, although we do not restrict communication to speech acts. As a special case, we sometimes consider a situation where $a = b$. When a communicates a sentence σ to b, we assume that any information that is logically implied by the sentence is implicitly communicated. For instance, if one communicates the sentence $p\wedge q$, then one also implicitly communicates p and implicitly communicates q. We also assume that communication is *instantaneous* such that a hearer recognizes information at the moment when it is dispatched from a speaker. For instance, if a speaker a tells a sentence σ to a hearer b in a conversation, information in σ is conveyed to the hearer at the moment when the speaker utters the sentence.

The semantics of **BIC** is given by the Kripke semantics for normal modal operators. Formally, a **BIC** *model* is a tuple $(W, \nu, (\mathcal{B}_a)_{a\in A}, (\mathcal{I}_a)_{a\in A}, (\mathcal{C}_{ab})_{a,b\in A})$ where

- W is a set of possible worlds.
- $\nu : P \to 2^W$ is a truth assignment mapping each propositional constant to the set of worlds in which it is true.

[3] By an agent we mean a human agent or an artificial agent like a robot or a computer program.

- $\mathcal{B}_a \subseteq W \times W$ is a serial, transitive and Euclidean binary relation on W. That is, $\forall u \in W\, \exists v \in W,\ u\mathcal{B}_a v$; $\forall u, v, w \in W$, $u\mathcal{B}_a v$ and $v\mathcal{B}_a w$ imply $u\mathcal{B}_a w$; and $\forall u, v, w \in W$, $u\mathcal{B}_a v$ and $u\mathcal{B}_a w$ imply $v\mathcal{B}_a w$.
- $\mathcal{I}_a \subseteq W \times W$ is a serial relation on W, namely, $\forall u \in W\, \exists v \in W,\ u\mathcal{I}_a v$.
- $\mathcal{C}_{ab} \subseteq W \times W$ is a serial relation on W, namely, $\forall u \in W\, \exists v \in W,\ u\mathcal{C}_{ab} v$.
- $\mathcal{I}_a$ is transitive and Euclidean over $\mathcal{B}_a$. That is, $\forall u, v, w \in W$, $u\mathcal{B}_a v$ and $v\mathcal{I}_a w$ imply $u\mathcal{I}_a w$; and $\forall u, v, w \in W$, $u\mathcal{B}_a v$ and $u\mathcal{I}_a w$ imply $v\mathcal{I}_a w$.
- $\mathcal{C}_{ab}$ is transitive and Euclidean over $\mathcal{B}_a$. That is, $\forall u, v, w \in W$, $u\mathcal{B}_a v$ and $v\mathcal{C}_{ab} w$ imply $u\mathcal{C}_{ab} w$; and $\forall u, v, w \in W$, $u\mathcal{B}_a v$ and $u\mathcal{C}_{ab} w$ imply $v\mathcal{C}_{ab} w$.
- $\mathcal{C}_{ab}$ is transitive and Euclidean over $\mathcal{I}_a$. That is, $\forall u, v, w \in W$, $u\mathcal{I}_a v$ and $v\mathcal{C}_{ab} w$ imply $u\mathcal{C}_{ab} w$; and $\forall u, v, w \in W$, $u\mathcal{I}_a v$ and $u\mathcal{C}_{ab} w$ imply $v\mathcal{C}_{ab} w$.

Intuitively, $u\mathcal{B}_a v$ (resp. $u\mathcal{I}_a v$) means that v is compatible with a's belief (resp. intention) at u. Likewise, $u\mathcal{C}_{ab} v$ means that v is compatible with u where a's communication to b has taken place. The satisfaction of a sentence in a possible world $w \in W$ of a **BIC** model M is defined as follows.

- $(M, w) \models p$ iff $w \in \nu(p)$ for $p \in P$.
- $(M, w) \models \neg\varphi$ iff $(M, w) \not\models \varphi$.
- $(M, w) \models \varphi \wedge \psi$ iff $(M, w) \models \varphi$ and $(M, w) \models \psi$.
- $(M, w) \models B_a\varphi$ iff $(M, u) \models \varphi$ for every $u \in W$ s.t. $w\mathcal{B}_a u$.
- $(M, w) \models I_a\varphi$ iff $(M, u) \models \varphi$ for every $u \in W$ s.t. $w\mathcal{I}_a u$.
- $(M, w) \models C_{ab}\varphi$ iff $(M, u) \models \varphi$ for every $u \in W$ s.t. $w\mathcal{C}_{ab} u$.

A sentence φ is *true in a model M at a world w* iff $(M, w) \models \varphi$. φ is *true in a model M* (written $M \models \varphi$) iff $(M, w) \models \varphi$ for any $w \in W$. φ is *valid* iff $M \models \varphi$ for any **BIC** model M.

The logic **BIC** has the following axioms and inference rules:

1. **(P)** All propositional tautologies.
2. The axioms for B (the system KD45):

   ($\mathbf{K_B}$) $B_a\varphi \wedge B_a(\varphi \supset \psi) \supset B_a\psi$

   ($\mathbf{D_B}$) $\neg B_a\bot$

   ($\mathbf{4_B}$) $B_a\varphi \supset B_a B_a\varphi$

   ($\mathbf{5_B}$) $\neg B_a\varphi \supset B_a\neg B_a\varphi$
3. The axioms for I and C (the system KD):

   ($\mathbf{K_I}$) $I_a\varphi \wedge I_a(\varphi \supset \psi) \supset I_a\psi$

($\mathbf{D_I}$) $\neg I_a \bot$

($\mathbf{K_C}$) $C_{ab}\varphi \wedge C_{ab}(\varphi \supset \psi) \supset C_{ab}\psi$

($\mathbf{D_C}$) $\neg C_{ab}\bot$

4. The bridge axioms among B, I and C:

($\mathbf{4_{IB}}$) $I_a\varphi \supset B_a I_a\varphi$

($\mathbf{5_{IB}}$) $\neg I_a\varphi \supset B_a \neg I_a\varphi$

($\mathbf{4_{CB}}$) $C_{ab}\varphi \supset B_a C_{ab}\varphi$

($\mathbf{5_{CB}}$) $\neg C_{ab}\varphi \supset B_a \neg C_{ab}\varphi$

($\mathbf{4_{CI}}$) $C_{ab}\varphi \supset I_a C_{ab}\varphi$

($\mathbf{5_{CI}}$) $\neg C_{ab}\varphi \supset I_a \neg C_{ab}\varphi$

5. Rules of inference:

(MP) From φ and $\varphi \supset \psi$ infer ψ.

($\mathbf{N_B}$) From φ infer $B_a\varphi$.

($\mathbf{N_I}$) From φ infer $I_a\varphi$.

($\mathbf{N_C}$) From φ infer $C_{ab}\varphi$.

The following inference rules are derived in **BIC**.

($\mathbf{E_B}$) From $\varphi \equiv \psi$ infer $B_a\varphi \equiv B_a\psi$.

($\mathbf{E_I}$) From $\varphi \equiv \psi$ infer $I_a\varphi \equiv I_a\psi$.

($\mathbf{E_C}$) From $\varphi \equiv \psi$ infer $C_{ab}\varphi \equiv C_{ab}\psi$.

The definition of **BIC** theorems is standard: a **BIC** theorem is a formula that is obtained from axiom instances via the inference rules. We write $\vdash \varphi$ iff a sentence φ is a theorem of **BIC**. Some useful theorems are listed below [18]:

($\mathbf{D_\Box}$) $\vdash \Box\varphi \supset \neg\Box\neg\varphi$

($\mathbf{R_\Box}$) $\vdash \Box\varphi \wedge \Box\psi \equiv \Box(\varphi \wedge \psi)$

($\mathbf{C_\Box}$) $\vdash \Box\varphi \vee \Box\psi \supset \Box(\varphi \vee \psi)$

where $\Box$ is either B_a, I_a or C_{ab}. It is well-known that: $\vdash (\Box\varphi \supset \neg\Box\neg\varphi) \equiv \neg\Box\bot$. When $\Box$ is B_a (resp. I_a or C_{ab}), the theorem $\mathbf{D}_\Box$ is referred to as $\mathbf{D_B}$ (resp. $\mathbf{D_I}$ or $\mathbf{D_C}$). Similar reference will be done for $\mathbf{R}_\Box$ and $\mathbf{C}_\Box$.

Proposition 2.1 *Let $\Box_i$ $(i = 1, 2)$ be either B_a, I_a or C_{ab}.*

$$\vdash (\,\Box_1\Box_2\varphi \wedge \Box_1\Box_2\psi\,) \equiv \Box_1\Box_2(\varphi \wedge \psi).$$

Proof: By $(\mathbf{E_B})$ and $(\mathbf{R}_\Box)$, it holds that $\vdash B_a(\Box\varphi \wedge \Box\psi) \equiv B_a\Box(\varphi \wedge \psi)$. By $(\mathbf{R_B})$, it holds that $\vdash B_a(\Box\varphi \wedge \Box\psi) \equiv B_a\Box\varphi \wedge B_a\Box\psi$. Hence, $\vdash B_a\Box\varphi \wedge B_a\Box\psi \equiv B_a\Box(\varphi \wedge \psi)$. Similarly, we can show that: $\vdash I_a\Box\varphi \wedge I_a\Box\psi \equiv I_a\Box(\varphi \wedge \psi)$ and $\vdash C_{ab}\Box\varphi \wedge C_{ab}\Box\psi \equiv C_{ab}\Box(\varphi \wedge \psi)$. Hence, the result holds. □

The result of Proposition 2.1 is easily extended to:

$$\vdash (\, \Box_1\Box_2 \cdots \Box_k\varphi \wedge \Box_1\Box_2 \cdots \Box_k\psi \,) \equiv \Box_1\Box_2 \cdots \Box_k(\varphi \wedge \psi)$$

by repeatedly applying $(\mathbf{E}_\Box)$ and $(\mathbf{R}_\Box)$.

One may argue that the axiom $(\mathbf{K_C})$ is too strong because if a communicates φ to b, all that φ deductively entails is communicated as well. A similar problem happens with $(\mathbf{K_B})$, however, one believes every theorem—a well-known *logical omniscience* [29]. This is a fundamental difficulty that lies in the nature of Kripke-style modal semantics for modelling knowledge and belief, but we do not address the issue further in this paper. By $(\mathbf{N_B})$ and $(\mathbf{K_B})$, each agent believes that other agents follow the same logic as itself. Thus, "$B_aB_b\varphi \supset B_a\neg B_b\neg\varphi$" and "$B_a(I_b\varphi \wedge I_b(\varphi \supset \psi)) \supset B_aI_b\psi$" are **BIC** theorems, for instance. The necessitation rule $(\mathbf{N_I})$ says that every theorem holds at all states of affairs that an agent a might intend to bring about. $(\mathbf{N_C})$ says that every theorem is unconditionally communicated from a to b. This is the case since every agent shares theorems under the same logic. Technically, both $(\mathbf{N_I})$ and $(\mathbf{N_C})$ are needed for the completeness with respect to Kripke models.[4]

Each axiom has its semantic correspondence as follows [74]. The axioms $(\mathbf{D_B})$, $(\mathbf{D_I})$, and $(\mathbf{D_C})$ respectively correspond to seriality of the relations $\mathcal{B}_a$, $\mathcal{I}_a$, and $\mathcal{C}_{ab}$. The axioms $(\mathbf{4_B})$ and $(\mathbf{5_B})$ respectively correspond to transitivity and Euclidianity of the relation $\mathcal{B}_a$. The axiom $(\mathbf{4_{IB}})$ corresponds to transitivity of $\mathcal{I}_a$ over $\mathcal{B}_a$, and $(\mathbf{5_{IB}})$ corresponds to Euclidianity of $\mathcal{I}_a$ over $\mathcal{B}_a$. The axioms $(\mathbf{4_{CB}})$, $(\mathbf{5_{CB}})$, $(\mathbf{4_{CI}})$, and $(\mathbf{5_{CI}})$ respectively correspond to their counterparts in similar ways. With this correspondence, it is easy to see that the axiomatic system of **BIC** is sound. Moreover, these axioms characterize *canonical* models [74], so that the axiomatic system is also complete.

2.2 Sincere communication

In this paper we consider communication which satisfies the following two conditions:

- A statement is made by a speaker (**statement condition**).
- A statement is received by a hearer (**addressee condition**).

The statement condition says that communication accompanies a statement of a sentence. The addressee condition says that communication requires the existence of a hearer. Communication is *sincere* if it satisfies the condition:

- A speaker believes the statement to be true (**truthfulness condition**).

The truthfulness condition represents the *sincerity rule* of the speech-act theory. Moreover, sincere communication is *intentional* if it satisfies the condition:

[4] Similar usages are observed in [23, 40, 48, 73].

- A speaker intends that a hearer believes the statement to be true (**intention condition**).

The condition says that intentional sincere communication is not simply saying something that one believes to be true, but involves an intention to make a hearer believe it. Note that as remarked in Section 2.1 we assume communication such that a hearer recognizes a statement *at the same moment* when the speaker makes the statement. Thus, intentional sincere communication is such that a speaker believes a sentence and communicates it to a hearer, while intending the hearer's believing the sentence at the moment when the communication is taken place.[5]

Under the logic **BIC**, (intentional) sincere communication is formulated as follows.

Definition 2.1 (sincere communication) Let a and b be two agents and $\sigma \in \Phi$. Then a *sincere communication* is defined as follows.

$$SINC_{ab}(\sigma) \stackrel{def}{=} C_{ab}\sigma \wedge B_a\sigma. \tag{1}$$

In this case, we say that an agent a *sincerely communicates* a sentence σ to another agent b. By contrast, an *intentional sincere communication* is defined as follows.

$$I\text{-}SINC_{ab}(\sigma) \stackrel{def}{=} SINC_{ab}(\sigma) \wedge I_a B_b \sigma. \tag{2}$$

By definition, communication from an agent a to another agent b is sincere if a communicates a believed-true sentence σ to b. In (1), both the statement condition and the addressee condition are represented by $C_{ab}\sigma$, and the truthfulness condition is represented by $B_a\sigma$. In (2), the additional condition $I_a B_b \sigma$ is imposed on $SINC_{ab}(\sigma)$, which represents that a has the intention that σ is believed by b. In what follows, $(I\text{-})SINC_{ab}(\sigma)$ means either $SINC_{ab}(\sigma)$ or $I\text{-}SINC_{ab}(\sigma)$. (Intentional) sincere communication has the following properties.

Proposition 2.2 (sincere communication on valid or contradictory sentences)

1. $\vdash (I\text{-})SINC_{ab}(\top)$.
2. $\vdash (I\text{-})SINC_{ab}(\bot) \supset \bot$.

Proof: 1. The result follows by the fact that $C_{ab}\top$, $B_a\top$, and $I_a B_b \top$ are all theorems of **BIC**. *2.* $(I\text{-})SINC_{ab}(\bot)$ implies $C_{ab}\bot$ that contradicts ($\mathbf{D_C}$). □

Proposition 2.3 (sincere communication on combined sentences)

1. $\vdash (\, SINC_{ab}(\lambda) \wedge SINC_{ab}(\sigma)\,) \equiv SINC_{ab}(\lambda \wedge \sigma)$.
2. $\vdash (\, I\text{-}SINC_{ab}(\lambda) \wedge I\text{-}SINC_{ab}(\sigma)\,) \equiv I\text{-}SINC_{ab}(\lambda \wedge \sigma)$.

[5] To characterize belief change of a hearer, temporal or dynamic logic would be more appropriate. However, we do not formulate the effect of communication nor belief change of a hearer. Whether or not a speaker makes sincere communication depends only on the belief and intention of a speaker, and is independent of the effect of the action. This is in contrast to *deception* discussed in Section 6.2.

Proof: *1.* $SINC_{ab}(\lambda) \wedge SINC_{ab}(\sigma) \equiv (C_{ab}\lambda \wedge C_{ab}\sigma) \wedge (B_a\lambda \wedge B_a\sigma)$. By two theorems ($\mathbf{R_C}$) and ($\mathbf{R_B}$), it holds that $\vdash C_{ab}\lambda \wedge C_{ab}\sigma \equiv C_{ab}(\lambda \wedge \sigma)$ and $\vdash B_a\lambda \wedge B_a\sigma \equiv B_a(\lambda \wedge \sigma)$. Hence, the result holds.
2. $I\text{-}SINC_{ab}(\lambda) \wedge I\text{-}SINC_{ab}(\sigma) \equiv SINC_{ab}(\lambda \wedge \sigma) \wedge I_aB_b\lambda \wedge I_aB_b\sigma$. By Proposition 2.1, $\vdash I_aB_b\lambda \wedge I_aB_b\sigma \equiv I_aB_b(\lambda \wedge \sigma)$ holds. Hence, the result holds. □

Proposition 2.4 (sincere communication on contrary sentences)

1. $\vdash (\, SINC_{ab}(\sigma) \wedge SINC_{ab}(\neg\sigma)\,) \supset \bot$.

2. $\vdash (\, I\text{-}SINC_{ab}(\sigma) \wedge I\text{-}SINC_{ab}(\neg\sigma)\,) \supset \bot$.

Proof: The results hold by Propositions 2.2 and 2.3. □

A sentence of the form $\sigma \wedge \neg B_a\sigma$ is known as *Moore's paradoxical sentence* [47]. The Moore sentence is not communicable in sincere communication.

Proposition 2.5 (sincere communication on Moore sentence)

$$\vdash (I\text{-})SINC_{ab}(\sigma \wedge \neg B_a\sigma) \supset \bot\,.$$

Proof: $(I\text{-})SINC_{ab}(\sigma \wedge \neg B_a\sigma) \equiv (I\text{-})SINC_{ab}(\sigma) \wedge (I\text{-})SINC_{ab}(\neg B_a\sigma)$ by Proposition 2.3. $(I\text{-})SINC_{ab}(\sigma) \wedge (I\text{-})SINC_{ab}(\neg B_a\sigma)$ implies $B_a\sigma \wedge B_a\neg B_a\sigma$. $B_a\sigma$ implies $B_aB_a\sigma$ ($\mathbf{4_B}$), while $B_a\neg B_a\sigma$ implies $\neg B_aB_a\sigma$ ($\mathbf{D_B}$). Contradiction. □

In speech act theory it is often assumed that communicating a sentence σ counts as an expression of the speaker's belief that σ.[6] This assumption is formally represented as follows.

Definition 2.2 (communicating belief) *Let* a *and* b *be two agents and* $\sigma \in \Phi$.

$$Com\_Bel_{ab}(\sigma) \stackrel{def}{=} C_{ab}\sigma \supset C_{ab}B_a\sigma. \tag{3}$$

With this assumption, the following result holds.

Proposition 2.6 (sincere communication on belief)

$$\vdash (SINC_{ab}(\sigma) \wedge Com\_Bel_{ab}(\sigma)) \supset SINC_{ab}(B_a\sigma)\,.$$

Proof: $SINC_{ab}(\sigma)$ implies $C_{ab}\sigma \wedge B_a\sigma$. $C_{ab}\sigma \wedge Com\_Bel_{ab}(\sigma)$ implies $C_{ab}B_a\sigma$. $B_a\sigma$ implies $B_aB_a\sigma$ by ($\mathbf{4_B}$). Hence, the result holds. □

An agent has a belief of his/her communication.

Proposition 2.7 (introspection)

$$\vdash (I\text{-})SINC_{ab}(\sigma) \supset B_a\,(I\text{-})SINC_{ab}(\sigma)\,.$$

Proof: The result holds by Definition 2.1 and the axioms ($\mathbf{4_{CB}}$), ($\mathbf{4_B}$), and ($\mathbf{4_{IB}}$). □

[6] "Thus to assert, affirm, state (that *p*) counts as an *expression of belief* (that *p*)" [62, p. 65].

Communication involves a speaker and a hearer. On the other hand, we can consider the special case of communicating to oneself, for instance, taking a note for remembering or keeping a diary. Normally, if one believes a fact, then he/she does not intend to make oneself disbelieve the fact. We call it a *rational balance* between belief and intention.[7] Such a rational balance between belief and intention is formally represented as follows.

Definition 2.3 (rational balance) Let a be an agent and $\sigma \in \Phi$.

$$RB_a(\sigma) \stackrel{def}{=} B_a\sigma \supset \neg I_a \neg B_a\sigma\,. \tag{4}$$

In this case, we say that a is *rationally balanced* on the sentence σ.

Proposition 2.8 (intentional sincere communication is rationally balanced)

$$\vdash I\text{-}SINC_{aa}(\sigma) \supset RB_a(\sigma)\,.$$

Proof: $I\text{-}SINC_{aa}(\sigma)$ implies $I_aB_a\sigma$, which implies $\neg I_a\neg B_a\sigma$ by ($\mathbf{D_I}$). □

3 Lies

In this section, we first provide two different definitions of *lies* based on Mahon [42] and investigate their formal properties. We then consider incentives behind the act of lying[8] and introduce the notion of *lies with objectives*.

3.1 Two definitions of lies

A *lie* is a representative of dishonest acts by people. According to Wikipedia, there are 30 different types of lies.[9] In spite of its familiarity to most of us, the question of "What is lying?" has been subject to extensive studies by a number of philosophers [5, 7, 14, 30, 42, 64]. Mahon [43] argues that the most common definition of lies requires the statement condition, the addressee condition and the intention condition of Section 2.2. On the other hand, instead of truthfulness of sincere communication, lies require the condition:

- A speaker believes the statement to be false (**untruthfulness condition**).

Note that the untruthfulness condition says that a speaker believes the falsity of the statement but the actual falsity of the statement is not requested. So if a speaker makes a believed-false statement which is in fact true, then the speaker's act is considered a

[7] The term is borrowed from [21]. Note that the intention here is the *present-directed* intention which is an intention to do some action now, and is different from the *future-directed* one which is an intention to do some action later [9].

[8] Some researcher distinguishes "lying" and "telling a lie" in the literature [64], while we do not distinguish *lying* and *lies* and use those terms interchangeably in this paper.

[9] Lie. In *Wikipedia: The Free Encyclopedia*. Retrieved November 2014, from http://en.wikipedia.org/wiki/Lie

lie.[10] Note also that the untruthfulness condition focuses on the content of the statement, so if a statement is made not by a speaker but in fact by some impostor, the statement is not considered untruthful as far as it contains no information of the speaker. Moreover, lies involve an intention to deceive on the statement made by a speaker. Thus, if one says something manifestly false as a joke or a metaphor, it is not a lie.[11]

Among a number of different definitions of lies, Mahon [42] states that the four necessary conditions are actually sufficient.

> *To lie (to another person) is: to make a believed-false statement (to another person) with the intention that that statement be believed to be true (by the other person)*—[38] and [42, (L6)].

The definition is formulated in **BIC** as follows.

Definition 3.1 (lie) Let a and b be two agents and $\sigma \in \Phi$.

$$LIE_{ab}(\sigma) \stackrel{def}{=} C_{ab}\sigma \wedge B_a\neg\sigma \wedge I_aB_b\sigma. \tag{5}$$

In this case, we say that a *lies* to b on the sentence σ. σ is also called a *lie*.

By definition, a lies to b if a communicates a believed-false sentence σ to b with the intention that σ is believed by b. (5) satisfies both the statement condition and the addressee condition ($C_{ab}\sigma$). (5) also satisfies the untruthfulness condition ($B_a\neg\sigma$) and the intention condition ($I_aB_b\sigma$).[12]

Mahon also argues that although the above definition seems to be suitable for most purposes, it does have the relatively strong requirement that a speaker intends to deceive on the direct contents of his/her statement. However, a speaker may very well have an intention to deceive about his/her belief in the truth of the statement that he/she makes, which we call an *intention to deceive about truthfulness*.[13] By contrast, we also call the condition $I_aB_b\sigma$ of (5) an *intention to deceive about truth*. Borrowing an example of [42], suppose that an FBI agent is working undercover in a criminal organization. The crime boss notices this fact, but the FBI agent has no suspicion of this. If the crime boss tells the FBI agent that there are no informants in his organization, then the boss cannot intend that the FBI agent believes this statement to be true because the boss knows that the agent is an informant. In this case, the crime boss can only intend that the FBI agent believes that the boss believes this statement to be true. According to the previous definition, the boss is not lying to the FBI agent. To cope with such cases, Mahon proposes another definition as follows:

> *To lie (to another person) is: to make a believed-false statement (to another person), either with the intention that that statement be believed to be true (by the other person), or with the intention that it be believed (by*

[10] "A person is to be judged as lying or not lying according to the intention of his own mind, not according to the truth or falsity of the matter itself" [5, p.55].

[11] Some philosophers argue that an intention to deceive is not a necessary condition of lying, however [14].

[12] Note again that a speaker believes $\neg\sigma$ while communicating σ to a hearer by intending the hearer's believing σ at the moment when the communication is taken place.

[13] In [43], it is called the *believed truthfulness condition*.

the other person) that that statement is believed to be true (by the person making the statement), or with both intentions—[42, (L6*)].

Lies about truthfulness are formulated in **BIC** as follows.

Definition 3.2 (lie about truthfulness) Let a and b be two agents and $\sigma \in \Phi$.

$$LIE^B_{ab}(\sigma) \stackrel{def}{=} C_{ab}\sigma \wedge B_a\neg\sigma \wedge I_aB_bB_a\sigma. \tag{6}$$

In (6), an intention to deceive about truthfulness is represented by $I_aB_bB_a\sigma$. In contrast to Definition 3.2, lies in Definition 3.1 are called *lies about truth*. Using (6), Mahon's second definition (L6*) of lies is stated in **BIC** as follows.

$$LIE^*_{ab}(\sigma) \stackrel{def}{=} LIE_{ab}(\sigma) \vee LIE^B_{ab}(\sigma)\,. \tag{7}$$

Mahon contends that these two definitions, (L6) and (L6*) of [42], are the best definitions of lies. In what follows, $LIE^{(B)}_{ab}(\sigma)$ means either $LIE_{ab}(\sigma)$ or $LIE^B_{ab}(\sigma)$. As the FBI example shows, "intention to deceive about truthfulness" might be the case without "intention to deceive about truth". The relation between lies about truth and lies about truthfulness is characterized using (3) as follows.

Proposition 3.1 (relation between LIE and LIE^B)

$$\vdash (LIE^B_{ab}(\sigma) \wedge Com\_Bel_{ab}(\sigma)) \supset LIE_{ab}(B_a\sigma).$$

Proof: $LIE^B_{ab}(\sigma)\wedge Com\_Bel_{ab}(\sigma)$ implies $C_{ab}B_a\sigma\wedge B_a\neg\sigma\wedge I_aB_bB_a\sigma$. Since $B_a\neg\sigma$ implies $B_a\neg B_a\sigma$ by ($\mathbf{D_B}$) and ($\mathbf{5_B}$), the result holds. □

Lies are not sincere communication.

Proposition 3.2 (lie is not sincere)

$$\vdash (LIE^{(B)}_{ab}(\sigma) \wedge (I\text{-})SINC_{ab}(\sigma)) \supset \bot.$$

Proof: $LIE^{(B)}_{ab}(\sigma) \wedge (I\text{-})SINC_{ab}(\sigma)$ implies $B_a\neg\sigma \wedge B_a\sigma$, which implies $\neg B_a\sigma \wedge B_a\sigma$ ($\mathbf{D_B}$). Contradiction. □

A lie on valid or contradictory sentences is meaningless.

Proposition 3.3 (lie on valid or contradictory sentences)

1. $\vdash LIE^{(B)}_{ab}(\top) \supset \bot$.
2. $\vdash LIE^{(B)}_{ab}(\bot) \supset \bot$.

Proof: 1. Both $LIE_{ab}(\top)$ and $LIE^B_{ab}(\top)$ imply $B_a\bot$ that contradicts ($\mathbf{D_B}$).
2. Both $LIE_{ab}(\bot)$ and $LIE^B_{ab}(\bot)$ imply $C_{ab}\bot$ that contradicts ($\mathbf{D_C}$). □

Lies on two sentences imply a lie on their conjunction.

Proposition 3.4 (lie on combined sentences)

1. $\vdash (\, LIE_{ab}(\lambda) \wedge LIE_{ab}(\sigma)\,) \supset LIE_{ab}(\lambda \wedge \sigma)$.
2. $\vdash (\, LIE^B_{ab}(\lambda) \wedge LIE^B_{ab}(\sigma)\,) \supset LIE^B_{ab}(\lambda \wedge \sigma)$.

Proof: 1. $LIE_{ab}(\lambda) \wedge LIE_{ab}(\sigma) \equiv C_{ab}\lambda \wedge C_{ab}\sigma \wedge B_a\neg\lambda \wedge B_a\neg\sigma \wedge I_aB_b\lambda \wedge I_aB_b\sigma$. By $(\mathbf{R_C})$, $(\mathbf{R_B})$ and Proposition 2.1, it is equivalent to $C_{ab}(\lambda \wedge \sigma) \wedge B_a(\neg\lambda \wedge \neg\sigma) \wedge I_aB_b(\lambda \wedge \sigma)$. Since $\vdash B_a(\neg\lambda \wedge \neg\sigma) \supset B_a\neg(\lambda \wedge \sigma)$ by $(\mathbf{P})$, $(\mathbf{N_B})$ and $(\mathbf{K_B})$, the result holds.
2. By Proposition 2.1, $B_bB_a\lambda \wedge B_bB_a\sigma \equiv B_bB_a(\lambda \wedge \sigma)$. By $(\mathbf{E_I})$ and $(\mathbf{R_I})$, we have $I_aB_bB_a\lambda \wedge I_aB_bB_a\sigma \equiv I_aB_bB_a(\lambda \wedge \sigma)$. Hence, the result holds by the proof of 1. □

The converse implication of Proposition 3.4 does not hold in general. For instance, if a job applicant says he/she has skill in both typing and technical writing and he/she in fact has no skill in technical writing, he/she is lying. But this does not imply that he/she is lying on the skill in typing.

A single lie makes the whole communication a lie.

Proposition 3.5 (combining lie and sincere communication)

$$\vdash (\, LIE_{ab}(\lambda) \wedge I\text{-}SINC_{ab}(\sigma)\,) \supset LIE_{ab}(\lambda \wedge \sigma).$$

Proof: $LIE_{ab}(\lambda) \wedge I\text{-}SINC_{ab}(\sigma)$ implies $C_{ab}\lambda \wedge C_{ab}\sigma$, $B_a\neg\lambda \wedge B_a\sigma$, and $I_aB_b\lambda \wedge I_aB_b\sigma$. Since $\vdash B_a\neg\lambda \wedge B_a\sigma \equiv B_a(\neg\lambda \wedge \sigma)$ by $(\mathbf{R_B})$ and $\vdash B_a(\neg\lambda \wedge \sigma) \supset B_a\neg(\lambda \wedge \sigma)$ by $(\mathbf{P})$, $(\mathbf{N_B})$ and $(\mathbf{K_B})$, the result holds. □

Lies on contrary sentence are meaningless.

Proposition 3.6 (lie on contrary sentences)

1. $\vdash (\, LIE_{ab}(\sigma) \wedge LIE_{ab}(\neg\sigma)\,) \supset \bot$.
2. $\vdash (\, LIE^B_{ab}(\sigma) \wedge LIE^B_{ab}(\neg\sigma)\,) \supset \bot$.

Proof: The results hold by Proposition 3.3(2) and Proposition 3.4. □

When an agent lies about truthfulness, he/she cannot lie on Moore sentences.

Proposition 3.7 (lie on Moore sentence)

$$\vdash LIE^B_{ab}(\sigma \wedge \neg B_a\sigma) \supset \bot.$$

Proof: $LIE^B_{ab}(\sigma \wedge \neg B_a\sigma)$ implies $I_aB_bB_a(\sigma \wedge \neg B_a\sigma)$ which is equivalent to $I_aB_b(B_a\sigma \wedge B_a\neg B_a\sigma)$ (Proposition 2.1). Since $B_a\sigma$ implies $B_aB_a\sigma$ $(\mathbf{4_B})$ and $B_a\neg B_a\sigma$ implies $\neg B_aB_a\sigma$ $(\mathbf{D_B})$, $\vdash B_a\sigma \wedge B_a\neg B_a\sigma \supset \bot$. By $(\mathbf{N_B})$, $(\mathbf{K_B})$, $(\mathbf{N_I})$, and $(\mathbf{K_I})$, $\vdash I_aB_b(B_a\sigma \wedge B_a\neg B_a\sigma) \supset I_aB_b\bot$. Hence, $I_aB_b(B_a\sigma \wedge B_a\neg B_a\sigma)$ implies $I_aB_b\bot$. On the other hand, $\neg B_b\bot$ $(\mathbf{D_B})$ thereby $I_a\neg B_b\bot$ $(\mathbf{N_I})$ which implies $\neg I_aB_b\bot$ $(\mathbf{D_I})$. Contradiction. □

When an agent lies about truth, the next result holds.

Proposition 3.8 (lies on a sentence and its disbelief)

$$\vdash (LIE_{ab}(\sigma) \wedge LIE_{ab}(\neg B_a\sigma)) \supset \bot .$$

Proof: $LIE_{ab}(\sigma)$ implies $B_a\neg\sigma$, which implies $\neg B_a\sigma$ ($\mathbf{D_B}$). $\neg B_a\sigma$ implies $B_a\neg B_a\sigma$ ($\mathbf{5_B}$), which implies $\neg B_a B_a\sigma$ ($\mathbf{D_B}$). On the other hand, $LIE_{ab}(\neg B_a\sigma)$ implies $B_a B_a\sigma$. Contradiction. □

If an agent lies, he/she has a belief of his/her dishonest act.

Proposition 3.9 (introspection)

$$\vdash LIE_{ab}^{(B)}(\sigma) \supset B_a\, LIE_{ab}^{(B)}(\sigma) .$$

Proof: $LIE_{ab}(\sigma)$ and $LIE_{ab}^{B}(\sigma)$ respectively imply $B_a\, LIE_{ab}(\sigma)$ and $B_a\, LIE_{ab}^{B}(\sigma)$ by Defs 3.1 and 3.2 and the axioms ($\mathbf{4_{CB}}$), ($\mathbf{4_B}$), ($\mathbf{4_{IB}}$), and ($\mathbf{R_B}$). □

When an agent is rationally balanced, lying to oneself leads to contradiction.[14]

Proposition 3.10 (lie to oneself)

1. $\vdash (LIE_{aa}(\sigma) \wedge RB_a(\neg\sigma)) \supset \bot .$
2. $\vdash (LIE_{aa}^{B}(\sigma) \wedge RB_a(\neg B_a\sigma)) \supset \bot .$

Proof: 1. $LIE_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge I_a B_a\sigma$. By ($\mathbf{D_B}$), ($\mathbf{N_I}$) and ($\mathbf{K_I}$), it holds that $\vdash I_a B_a\sigma \supset I_a\neg B_a\neg\sigma$. So $I_a B_a\sigma$ implies $I_a\neg B_a\neg\sigma$. On the other hand, $RB_a(\neg\sigma) \wedge B_a\neg\sigma$ implies $\neg I_a\neg B_a\neg\sigma$. Contradiction.
2. $LIE_{aa}^{B}(\sigma)$ implies $B_a\neg\sigma \wedge I_a B_a B_a\sigma$, which implies $I_a\neg B_a\neg B_a\sigma$ by ($\mathbf{D_B}$), ($\mathbf{N_I}$), ($\mathbf{K_I}$) and (**MP**). Also, $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$) that implies $B_a\neg B_a\sigma$ ($\mathbf{5_B}$). By $RB_a(\neg B_a\sigma)$, $B_a\neg B_a\sigma$ implies $\neg I_a\neg B_a\neg B_a\sigma$, which contradicts $I_a\neg B_a\neg B_a\sigma$. □

By definition, lies depend only on the belief state of the speaker and his/her act of communication. In many cases, however, one has an incentive to lie. In the next subsection, we will consider conditions under which an agent decides to lie.

3.2 Lies with objectives

One usually has motives for lying and several reasons might be behind the act. Suppose that an agent has a desired outcome that he/she wants to obtain, while he/she believes that the outcome would not be achieved by telling true beliefs. On the other hand, he/she believes that the outcome might be achieved by telling false beliefs. In this case, the agent has an incentive to lie.

[14] "Self-deception exists, I will say, when a person lies to himself, that is to say, persuades himself to believe what he knows is not so. ... Thus, self-deception involves an inner conflict, perhaps the existence of contradiction" [25].

We call this *lies with objectives*. By lies with objectives, a speaker intends to lead a hearer to believe a particular sentence. An objective is an effect expected by a speaker with respect to the result of reasoning by a hearer. Thus, in lies with objectives a speaker reasons about what a hearer believes in the context of discourse. Lies with objectives are formally defined as follows.

Definition 3.3 (lie with objective) Let a and b be two agents and $\sigma, \varphi \in \Phi$.

$$O\text{-}LIE_{ab}(\sigma, \varphi) \stackrel{def}{=} I_a B_b\, \varphi \wedge \neg B_a B_b \varphi \wedge B_a B_b(\sigma \supset \varphi) \wedge B_a \neg\sigma \wedge C_{ab}\sigma\,. \qquad (8)$$

In this case, a *lies* to b on σ with an *objective* φ.

The intuitive meaning of (8) is as follows. An agent a lies on σ with an objective φ if (i) a has an intention to make b believe a sentence φ ($I_a B_b\, \varphi$), and (ii) a disbelieves that b believes φ ($\neg B_a B_b \varphi$), while (iii) a believes that the believed-false sentence σ leads b to believe φ ($B_a B_b(\sigma \supset \varphi) \wedge B_a \neg\sigma$), then (iv) a communicates σ to b ($C_{ab}\sigma$). Note that lies with objectives do not necessarily require a speaker a's intention of a hearer b's believing σ. The speaker intends to make the hearer believe an objective sentence φ, instead. For instance, when a woman gets a telephone call from a salesperson and says that she has something on the stove which is in fact false, she does not necessarily intend to make the salesperson believe that she is cooking but intends to make the salesperson believe that she is not able to talk now. As the case of lies, we can have the second definition for lies with objectives by replacing $B_a B_b(\sigma \supset \varphi)$ with $B_a B_b(B_a \sigma \supset \varphi)$.

Example 3.1 Suppose that a salesperson a is dealing with a customer b. The salesperson has the objective of the customer's buying a product. Then a has an intention to make b believe $\varphi = \texttt{buy}$, but disbelieves that b believes it:

$$I_a B_b\, \texttt{buy} \;\wedge\; \neg B_a B_b\, \texttt{buy}.$$

The salesperson also believes that the customer will buy the product if it has a high quality:

$$B_a B_b(\texttt{high\_quality} \supset \texttt{buy}).$$

The salesperson believes that the quality of the product is not high, but communicates the contrary to the customer:

$$B_a \neg \texttt{high\_quality} \wedge C_{ab}\, \texttt{high\_quality}.$$

In this case, $O\text{-}LIE_{ab}(\texttt{high\_quality}, \texttt{buy})$ holds.

A lie for the valid or contradictory objective sentence makes no sense.

Proposition 3.11 (lie with objective $\top$ or $\bot$)

1. $\vdash O\text{-}LIE_{ab}(\sigma, \top) \supset \bot$.
2. $\vdash O\text{-}LIE_{ab}(\sigma, \bot) \supset \bot$.

Proof: 1. $O\text{-}LIE_{ab}(\sigma, \top)$ implies $\neg B_a B_b \top$ by the second conjunct of (8), while $\vdash B_a B_b \top$ by a repeated application of ($\mathbf{N_B}$). Contradiction.
2. $O\text{-}LIE_{ab}(\sigma, \bot)$ implies $I_a B_b \bot$ by the first conjunct of (8), while from $\vdash \neg B_b \bot$ ($\mathbf{D_B}$) it holds that $\vdash I_a \neg B_b \bot$ ($\mathbf{N_I}$) thereby $\vdash \neg I_a B_b \bot$ ($\mathbf{D_I}$). Contradiction. □

As a special case, a may lie to b on the objective sentence φ to make b believe φ. The next result follows by the definition.

Proposition 3.12 (O-lie on the objective sentence)

$$O\text{-}LIE_{ab}(\varphi, \varphi) \equiv \neg B_a B_b \varphi \wedge LIE_{ab}(\varphi)\,.$$

In $O\text{-}LIE_{ab}(\varphi, \varphi)$, the condition $\neg B_a B_b \varphi$ means that a has motives for lying when a disbelieves that b believes the desired outcome φ. Thus, the definition of lies with objectives is stronger than the definition of lies of Definition 3.1. This is due to the fact that lies with objectives have an additional condition to have desired outcomes, so that if a believes that b already believes φ, there is no reason to lie anymore.

4 Bullshit, Withholding Information and Half-Truth

In this section, we formulate three different categories of dishonesty—*bullshit*, *withholding information* and *half-truth*. We investigate their formal properties and connections to lies.

4.1 Bullshit

Frankfurt [33] studies a category of dishonesty, called *bullshit*, that is different from lies. Bullshit is a statement that "is grounded neither in a belief that it is true nor, as a lie must be, in a belief that it is not true. It is just this lack of connection to a concern with truth—this indifference to how things really are—that I regarded as of the essence of bullshit" (ibid., pp. 33–34).[15] As an example, consider a financial consultant paid by the hour to provide advice to his clients. The consultant gives advice to buy stocks, for instance, but he may or may not believe that buying stocks is the best strategy (due to the lack of expertise). Bullshit is a quite common phenomenon in daily life. Frankfurt states a reason for its occurrence as follows: "Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about. Thus the production of bullshit is stimulated whenever a person's obligations or opportunities to speak about some topic exceed his knowledge of the facts that are relevant to that topic" (ibid., p.63). Bullshit can formally be defined in **BIC** as follows.

Definition 4.1 (bullshit) Let a and b be two agents and $\sigma \in \Phi$.

$$BS_{ab}(\sigma) \stackrel{def}{=} C_{ab}\sigma \wedge \neg B_a \sigma \wedge \neg B_a \neg \sigma. \tag{9}$$

In this case, we say that an agent a *bullshits* to another agent b on the sentence σ. σ is also called *bullshit* (for short, BS).

[15] Cohen [22] argues against this view and bullshit could be produced by a person who may concern about the truth. In this paper, however, we follow Frankfurt's definition.

In lies (Defs 3.1 and 3.2), the speaker a disbelieves σ but believes $\neg\sigma$. When bullshitting, on the other hand, a disbelieves $\neg\sigma$ as well. In other words, a has no belief with respect to the truth value of σ.[16]

Thus, BS satisfies the statement condition and the addressee condition of Section 2.2, while it does not satisfy the untruthfulness condition of Section 3.1. Moreover, BS does not satisfy the intention condition. In the example at the beginning of this section, the financial consultant has no interest in making the client believe that buying stocks is the best strategy or not. The only concern of the consultant is that the client believes that the statement is based on financial expertise. Since a speaker has no belief with respect to σ, there is a freedom for the speaker to communicate σ or $\neg\sigma$. The choice whether to communicate σ or $\neg\sigma$ depends on the speaker's opinion on how likely it will be for a hearer to believe one of them (given some additional explanation). This is in contrast to lies where speakers have no freedom to make this choice because one of these options (either σ or $\neg\sigma$) will have consequences they might want to enjoy. A liar usually has an interest in creating a particular belief at a hearer. This is not always the case for BS, however. Another difference is that while one can lie on one's own beliefs $LIE_{ab}(B_a\sigma)$, this is not the case for BS.

Proposition 4.1 (BS on one's own belief)

1. $\vdash BS_{ab}(B_a\sigma) \supset \bot$.

2. $\vdash BS_{ab}(\neg B_a\sigma) \supset \bot$.

Proof: Both $BS_{ab}(B_a\sigma)$ and $BS_{ab}(\neg B_a\sigma)$ imply $\neg B_a B_a\sigma \wedge \neg B_a \neg B_a\sigma$. Here $\neg B_a B_a\sigma$ implies $\neg B_a\sigma$ by contraposition of ($\mathbf{4_B}$), which implies $B_a \neg B_a\sigma$ ($\mathbf{5_B}$). This contradicts $\neg B_a \neg B_a\sigma$. □

Proposition 4.1 implies that BS on a sentence and its disbelief is impossible.

$$\vdash (\, BS_{ab}(\sigma) \wedge BS_{ab}(\neg B_a\sigma)\,) \supset \bot .$$

Like lies, BS is not sincere communication.

Proposition 4.2 (BS is not sincere)

$$\vdash (\, BS_{ab}(\sigma) \wedge SINC_{ab}(\sigma)\,) \supset \bot .$$

Proof: $BS_{ab}(\sigma) \wedge SINC_{ab}(\sigma)$ implies $\neg B_a\sigma \wedge B_a\sigma$. Contradiction. □

BS on valid or contradictory sentences is meaningless.

Proposition 4.3 (BS on valid or contradictory sentences)

1. $\vdash BS_{ab}(\top) \supset \bot$.

[16] Van Ditmarsch [76] provides a similar definition and calls it *bluff*. Frankfurt distinguishes them, however. Like lying, "bluffing, too, is typically devoted to conveying something false", on the other hand, "although it is produced without concern with the truth, it (bullshit) need not be false" [33, pp.46–47].

2. $\vdash BS_{ab}(\bot) \supset \bot$.

Proof: Both $BS_{ab}(\top)$ and $BS_{ab}(\bot)$ imply $\neg B_a \top$, but $\vdash B_a \top$ ($\mathbf{N_B}$). □

BS on contrary sentences is impossible.

Proposition 4.4 (BS on contrary sentences)

$$\vdash (\, BS_{ab}(\sigma) \wedge BS_{ab}(\neg\sigma)\,) \supset \bot\,.$$

Proof: $BS_{ab}(\sigma) \wedge BS_{ab}(\neg\sigma)$ implies $C_{ab}\sigma \wedge C_{ab}\neg\sigma$, so that $C_{ab}(\sigma \wedge \neg\sigma)$ by ($\mathbf{R_C}$). This contradicts ($\mathbf{D_C}$). □

In contrast to lies, BS on two sentences does not imply BS on their conjunction in general. For instance, suppose that a believes $\neg\lambda \vee \neg\sigma$, and bullshits to b on both λ and σ. In this case, $B_a(\neg\lambda \vee \neg\sigma) \wedge BS_{ab}(\lambda) \wedge BS_{ab}(\sigma)$ is consistent, while $B_a(\neg\lambda \vee \neg\sigma) \wedge BS_{ab}(\lambda \wedge \sigma)$ is inconsistent.

Combining BS and sincere communication does not produce sincere communication.

Proposition 4.5 (combining BS and sincere communication)

$$\vdash (\, BS_{ab}(\lambda) \wedge SINC_{ab}(\sigma)\,) \supset \neg SINC_{ab}(\lambda \wedge \sigma)\,.$$

Proof: $BS_{ab}(\lambda) \wedge SINC_{ab}(\sigma)$ implies $\neg B_a\lambda \wedge B_a\sigma$, while $SINC_{ab}(\lambda \wedge \sigma)$ implies $B_a(\lambda \wedge \sigma) \equiv B_a\lambda \wedge B_a\sigma$. Contradiction. □

BS satisfies the introspection condition.

Proposition 4.6 (introspection)

$$\vdash BS_{ab}(\sigma) \supset B_a\, BS_{ab}(\sigma)\,.$$

Proof: The result holds by Definition 4.1 and the axioms ($\mathbf{4_{CB}}$), ($\mathbf{5_B}$) and ($\mathbf{R_B}$). □

$BS_{ab}(\sigma)$ does not contradict the belief of the speaker a. So one cannot lie and BS on the same sentence.

Proposition 4.7 (lie and BS on the same sentence)

$$\vdash (\, LIE_{ab}(\sigma) \wedge BS_{ab}(\sigma)\,) \supset \bot\,.$$

Proof: $LIE_{ab}(\sigma)$ implies $B_a\neg\sigma$, while $BS_{ab}(\sigma)$ implies $\neg B_a\neg\sigma$. Contradiction. □

Sometimes BS accompanies intention. For instance, suppose a salesperson who is paid on commission basis, but does not really know the products that he is selling. The salesperson would make the claim that a product has a high quality, without having any knowledge on this. This is also an example of BS. However, making a client believe that the product has a high quality is preferred to making the client believe that the product has a low quality. The situation here differs from that of the financial consultant mentioned at the beginning of this subsection (who is paid by the hour by the client, and hence has no intrinsic interest to advise to buy stocks or not). Such *intentional bullshit* is defined next.

Definition 4.2 (intentional bullshit) Let a and b be two agents and $\sigma \in \Phi$.

$$I\text{-}BS_{ab}(\sigma) \stackrel{def}{=} BS_{ab}(\sigma) \wedge I_a B_b \sigma. \tag{10}$$

In this case, we say that an agent a *intentionally bullshits* to another agent b on the sentence σ. σ is also called *intentional bullshit* (for short, intentional BS).

By contrast, $BS_{ab}(\sigma)$ without $I_a B_b \sigma$ is called *unintentional*. In contrast to unintentional BS, intentional BS satisfies the intention condition. In this paper, we will ignore the difference between BS and I-BS in cases where it is unimportant. $(I\text{-})BS_{ab}(\sigma)$ means either $BS_{ab}(\sigma)$ or $I\text{-}BS_{ab}(\sigma)$. Like lies, it is also possible to define BS with an intention to deceive about truthfulness by replacing the condition $I_a B_b \sigma$ of (10) with $I_a B_b B_a \sigma$. In contrast to lies, a rationally-balanced agent can (intentionally) BS to oneself, that is, $(I\text{-})BS_{aa}(\sigma)$ is consistent with both $RB_a(\sigma)$ and $RB_a(\neg\sigma)$.

Intentional BS is also exclusive with intentional sincere communication, and combining intentional BS and intentional sincere communication results in no sincere communication. Moreover, when there is intentional communication of a sentence, sincere communication, lies and intentional BS are exhaustive.

Proposition 4.8 (exhaustiveness)

$$\vdash (C_{ab}\sigma \wedge I_a B_b \sigma) \equiv I\text{-}SINC_{ab}(\sigma) \vee LIE_{ab}(\sigma) \vee I\text{-}BS_{ab}(\sigma)\,.$$

Proof: $(C_{ab}\sigma \wedge I_a B_b \sigma) \equiv (C_{ab}\sigma \wedge I_a B_b \sigma \wedge (B_a\sigma \vee \neg B_a\sigma)) \equiv I\text{-}SINC_{ab}(\sigma) \vee (C_{ab}\sigma \wedge I_a B_b \sigma \wedge \neg B_a\sigma)$. Moreover, $(C_{ab}\sigma \wedge I_a B_b \sigma \wedge \neg B_a\sigma) \equiv (C_{ab}\sigma \wedge I_a B_b \sigma \wedge \neg B_a\sigma \wedge (B_a\neg\sigma \vee \neg B_a\neg\sigma)) \equiv (C_{ab}\sigma \wedge I_a B_b \sigma \wedge \neg B_a\sigma \wedge B_a\neg\sigma) \vee I\text{-}BS(\sigma)$. By $(\mathbf{D_B})$, $(C_{ab}\sigma \wedge I_a B_b \sigma \wedge \neg B_a\sigma \wedge B_a\neg\sigma) \equiv (C_{ab}\sigma \wedge I_a B_b \sigma \wedge B_a\neg\sigma) \equiv LIE_{ab}(\sigma)$. Hence, the result holds. □

BS can accompany objectives. Then BS with objectives is defined as follows.

Definition 4.3 (BS with objective) Let a and b be two agents and $\sigma, \varphi \in \Phi$.

$$O\text{-}BS_{ab}(\sigma, \varphi) \stackrel{def}{=} I_a B_b\, \varphi \wedge \neg B_a B_b \varphi \wedge B_a B_b(\sigma \supset \varphi) \wedge (I\text{-})BS_{ab}(\sigma). \tag{11}$$

In this case, a *(intentionally) bullshits* to b on σ with an *objective* φ.

In (11), a (intentionally) bullshits on σ with an objective φ if (i) a has an intention to make b believe a sentence φ, and (ii) a disbelieves that b believes φ, while (iii) a believes that the unknown sentence σ leads b to believe φ, then (iv) a (intentionally) bullshits to b on σ. As in the case of lies with objectives, the intention condition on the sentence σ is not necessarily needed. In particular, O-BS on the objective sentence becomes $O\text{-}BS_{ab}(\varphi, \varphi) \equiv \neg B_a B_b \varphi \wedge I\text{-}BS_{ab}(\varphi)$.

4.2 Withholding information

Sometimes the act of simply remaining silent with a deceptive intention is called "lie of omission" [42] or "withholding information" [15]. According to [15, p.56], "to withhold information is to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs." For instance, a job applicant who has a criminal record but does not inform the employer of this fact is withholding information. Withholding information is similar but different from *concealing* information. "To conceal information is to do things to hide information from someone—to prevent someone from discovering it" (ibid., p. 56). That is, a person just does not take an action of offering information in withholding information, while a person takes an action of hiding information in concealing information. It is worth noting that withholding or concealing information is not necessarily immoral. Suppose that a robber breaks into your home and demands to know where your valuables are. If you do not tell the location of valuables (with an intention that the robber does not believe it), then it would not be immoral. Withholding information is considered immoral "if there is a clear expectation, promise, and/or professional obligation that such information will be provided" (ibid, p. 56). Our current logic is not expressive enough to represent expectation, promise, or obligation. This section formulates withholding information in **BIC** for the purpose of contrasting it with lies or bullshit which is considered rather immoral.

We consider that an agent a withholds information σ from b when a makes no communication to b on a believed-true sentence σ with the intention that σ is disbelieved by b. Formally, it is defined as follows.

Definition 4.4 (withholding information) Let a and b be two agents and $\sigma \in \Phi$.

$$WI_{ab}(\sigma) \stackrel{def}{=} \neg C_{ab}\sigma \wedge B_a\sigma \wedge I_a\neg B_b\sigma. \tag{12}$$

In this case, we say that an agent a *withholds information* (for short, WI) σ from another agent b. σ is also called *withheld information*.

Unlike lies and BS, an agent makes no statement when withholding information. So the statement condition, the addressee condition, and the truthfulness condition of Section 2.2 are not satisfied, and the untruthfulness condition of Section 3.1 is also not satisfied. WI is not considered sincere communication in the sense that a speaker does not communicate a believed-true sentence.

Proposition 4.9 (WI is not sincere)

$$\vdash (WI_{ab}(\sigma) \wedge (I\text{-})SINC_{ab}(\sigma)) \supset \bot.$$

Proof: $WI_{ab}(\sigma) \wedge (I\text{-})SINC_{ab}(\sigma)$ implies $\neg C_{ab}\sigma \wedge C_{ab}\sigma$. Contradiction. □

Withholding valid or contradictory sentences is meaningless.

Proposition 4.10 (WI on valid or contradictory sentences)

1. $\vdash WI_{ab}(\top) \supset \bot$.

2. $\vdash WI_{ab}(\bot) \supset \bot$.

Proof: 1. $WI_{ab}(\top)$ implies $\neg C_{ab}\top$, however, $C_{ab}\top$ is a theorem by $(\mathbf{N_C})$.
2. $WI_{ab}(\bot)$ implies $B_a\bot$, which contradicts $(\mathbf{D_B})$. □

WI on two sentences implies WI on their conjunction.

Proposition 4.11 (WI on combined sentences)

$$\vdash (WI_{ab}(\lambda) \wedge WI_{ab}(\sigma)) \supset WI_{ab}(\lambda \wedge \sigma).$$

Proof: First, $WI_{ab}(\lambda) \wedge WI_{ab}(\sigma)$ implies $\neg C_{ab}\lambda \wedge \neg C_{ab}\sigma$, which implies $\neg C_{ab}\lambda \vee \neg C_{ab}\sigma$ by $(\mathbf{P})$, which is in turn equivalent to $\neg C_{ab}(\lambda \wedge \sigma)$ by $(\mathbf{R_C})$. Secondly, $WI_{ab}(\lambda) \wedge WI_{ab}(\sigma)$ implies $B_a\lambda \wedge B_a\sigma$, which is equivalent to $B_a(\lambda \wedge \sigma)$ by $(\mathbf{R_B})$. Thirdly, $WI_{ab}(\lambda) \wedge WI_{ab}(\sigma)$ implies $I_a\neg B_b\lambda \wedge I_a\neg B_b\sigma$ which is equivalent to $I_a(\neg B_b\lambda \wedge \neg B_b\sigma)$ by $(\mathbf{R_I})$, which implies $I_a(\neg B_b\lambda \vee \neg B_b\sigma)$ which is in turn equivalent to $I_a\neg B_b(\lambda \wedge \sigma)$ by $(\mathbf{R_B})$, $(\mathbf{P})$ and $(\mathbf{E_I})$. Hence, the result holds. □

Combining WI and sincere communication does not result in sincere communication.

Proposition 4.12 (combining WI and sincere communication)

$$\vdash (WI_{ab}(\lambda) \wedge (I\text{-})SINC_{ab}(\sigma)) \supset \neg (I\text{-})SINC_{ab}(\lambda \wedge \sigma).$$

Proof: $WI_{ab}(\lambda) \wedge (I\text{-})SINC_{ab}(\sigma)$ implies $\neg C_{ab}\lambda \wedge C_{ab}\sigma$, while $(I\text{-})SINC_{ab}(\lambda \wedge \sigma)$ implies $C_{ab}(\lambda \wedge \sigma)$ thereby $C_{ab}\lambda \wedge C_{ab}\sigma$ by $(\mathbf{R_C})$. Contradiction. □

Proposition 4.13 (WI on contrary sentences)

$$\vdash (WI_{ab}(\sigma) \wedge WI_{ab}(\neg\sigma)) \supset \bot.$$

Proof: $WI_{ab}(\sigma) \wedge WI_{ab}(\neg\sigma)$ implies $B_a\sigma \wedge B_a\neg\sigma$, contradiction. □

WI satisfies the introspection condition.

Proposition 4.14 (introspection)

$$\vdash WI_{ab}(\sigma) \supset B_a\, WI_{ab}(\sigma).$$

Proof: The result holds by Definition 4.4 and the axioms $(\mathbf{5_{CB}})$, $(\mathbf{4_B})$, and $(\mathbf{4_{IB}})$. □

When an agent is rationally balanced, withholding information from oneself is impossible.

Proposition 4.15 (WI from oneself)

$$\vdash (WI_{aa}(\sigma) \wedge RB_a(\sigma)) \supset \bot.$$

Proof: $WI_{aa}(\sigma)$ implies $B_a\sigma \wedge I_a\neg B_a\sigma$. By $RB_a(\sigma)$, $B_a\sigma$ implies $\neg I_a\neg B_a\sigma$. Contradiction. □

Lying on a sentence implies WI on its negated sentence.

Proposition 4.16 (lie implies WI)

$$\vdash LIE_{ab}(\sigma) \supset WI_{ab}(\neg\sigma)\,.$$

Proof: If $LIE_{ab}(\sigma)$ holds, then $C_{ab}\sigma \wedge B_a\neg\sigma \wedge I_aB_b\sigma$. $C_{ab}\sigma$ implies $\neg C_{ab}\neg\sigma$ ($\mathbf{D_C}$). $I_aB_b\sigma$ implies $I_a\neg B_b\neg\sigma$ by ($\mathbf{D_B}$), ($\mathbf{K_I}$), and ($\mathbf{N_I}$). Hence, the result holds. □

It is impossible to lie (or (I-)BS) and WI on the same sentence.

Proposition 4.17 (Lie, BS and WI on the same sentence)

$$\vdash (\,(LIE_{ab}(\sigma) \vee (I\text{-})BS_{ab}(\sigma)) \wedge WI_{ab}(\sigma)\,) \supset \bot\,.$$

Proof: $LIE_{ab}(\sigma)$ or $(I\text{-})BS_{ab}(\sigma)$ implies $C_{ab}\sigma$, but $WI_{ab}(\sigma)$ implies $\neg C_{ab}\sigma$. Contradiction. □

It is also possible to define WI with an intention to deceive about truthfulness by replacing $I_a\neg B_b\sigma$ with $I_a\neg B_bB_a\sigma$ in $WI_{ab}(\sigma)$. WI with objectives is defined as follows.

Definition 4.5 (WI with objective) Let a and b be two agents and $\sigma, \varphi \in \Phi$.

$$O\text{-}WI_{ab}(\sigma,\varphi) \stackrel{def}{=} I_aB_b\,\varphi \wedge \neg B_aB_b\varphi \wedge B_a(\neg B_b\sigma \supset B_b\varphi) \wedge WI_{ab}(\sigma). \qquad (13)$$

In this case, a *withholds information* σ from b with an *objective* φ.

In (13), a withholds information σ from b with an objective φ if (i) a has an intention to make b believe a sentence φ, and (ii) a disbelieves that b believes φ, while (iii) a believes that b's lacking information σ leads b to believe φ, then (iv) a withholds information σ from b. In particular, a can withhold $\neg\varphi$ to make b believe φ: $O\text{-}WI_{ab}(\neg\varphi,\varphi) \equiv I_aB_b\varphi \wedge \neg B_aB_b\varphi \wedge B_a(\neg B_b\neg\varphi \supset B_b\varphi) \wedge \neg C_{ab}\neg\varphi \wedge B_a\neg\varphi$. On the other hand, $O\text{-}WI_{ab}(\varphi,\varphi)$ implies $\neg B_aB_b\varphi \wedge B_a(\neg B_b\varphi \supset B_b\varphi)$ which is inconsistent.

4.3 Half-Truth

Half-truth is a partially true statement intended to deceive or mislead.[17] That is, a speaker makes a believed-true statement with the intention that a hearer misuses it to reach a wrong conclusion. It is often understood as *indirect lies* or *lying while saying the truth* [77]. For instance, John, who wants to marry his girlfriend Mary, tells her that he got a permanent position at a company. Mary then considers that John has a stable income now and would agree to marry him. The company is almost bankrupt,

[17] Collins English Dictionary.

however, and John believes that he would not get a stable income. But John does not tell Mary that his company is going bankrupt. In this speech act, John is telling the truth, while he expects that Mary will reach a conclusion "stable income" which he believes to be false. Thus, different from lies or BS, a speaker asserts what he/she believes true, while, at the same time, he/she conceals something of the truth hoping that a hearer will make an incorrect inference based on his/her belief.[18] Half-truths are formulated in **BIC** as follows.

Definition 4.6 (half-truth) Let a and b be two agents and δ, $\sigma \in \Phi$.

$$HT_{ab}(\sigma, \delta) \stackrel{def}{=} I\text{-}SINC_{ab}(\sigma) \wedge \neg B_a B_b \delta \wedge B_a B_b(\sigma \supset \delta) \\ \wedge B_a \neg\delta \wedge \neg C_{ab} \neg\delta \wedge I_a B_b \delta. \quad (14)$$

In this case, we say that an agent a provides a *half-truth* (for short, HT) to another agent b on the sentence σ to reach δ. σ is also called an HT.

The meaning of (14) is explained as follows. First, a communicates a believed-true sentence σ with the intention of making b believe it ($I\text{-}SINC_{ab}(\sigma)$). Secondly, a disbelieves that b believes δ ($\neg B_a B_b \delta$), while a believes that σ makes b believe δ. ($B_a B_b(\sigma \supset \delta)$). Thirdly, a believes the falsity of δ ($B_a \neg\delta$) but does not communicate $\neg\delta$ to b with the intention of making b believe δ ($\neg C_{ab} \neg\delta \wedge I_a B_b \delta$).

Compared with definitions of lies, BS and WI, one can observe that the act of HT is a bit complicated. In fact, "(t)he deceiver takes a more circuitous route to his success, where lying is an easier and more certain way to mislead" [1, p.440]. A reason for the complication is due to the fact that HT relies on the speaker's belief about the hearer's inference.[19] By definition, HT is an intentional sincere communication, so it satisfies the statement condition, the addressee condition, the truthfulness condition, and the intention condition of Section 2.2.

Example 4.1 The conversation between John and Mary in the above example is represented as follows. First, John (a) informs Mary (b) that he got a permanent position at a company with the intention that she believes the fact:

$$C_{ab}\,\texttt{permanent} \wedge B_a\,\texttt{permanent} \wedge I_a B_b\,\texttt{permanent}.$$

John disbelieves that Mary believes his stable income, but he believes that the permanent position makes Mary believe his stable income:

$$\neg B_a B_b \texttt{stable} \wedge B_a B_b(\texttt{permanent} \supset \texttt{stable}).$$

On the other hand, John believes that he will not get a stable income but does not inform Mary of this fact with the intention that she believes his stable income:

$$B_a \neg\texttt{stable} \wedge \neg C_{ab}\,\neg\,\texttt{stable} \wedge I_a B_b\,\texttt{stable}.$$

In this case, $HT_{ab}(\texttt{permanent}, \texttt{stable})$ holds.

[18] It is called *deception* in [1]. In this paper, however, we distinguish half-truths and deception that is explained later.

[19] An experimental study in neuroscience shows that increased demands on cognitive control arise when a person expresses the truth in an attempt to deceive another person [13].

One cannot tell a half-truth on σ to reach σ.

Proposition 4.18 (impossible HT)

$$\vdash HT_{ab}(\sigma, \sigma) \supset \bot .$$

Proof: $HT_{ab}(\sigma, \sigma)$ implies $B_a\sigma \wedge B_a\neg\sigma$, contradiction. □

HT on contradictory sentences is meaningless.

Proposition 4.19 (HT on contradictory sentences)

$$\vdash HT_{ab}(\bot, \delta) \supset \bot .$$

Proof: $HT_{ab}(\bot, \delta)$ implies $C_{ab}\bot$, which contradicts ($\mathbf{D_C}$). □

On the other hand, HT on valid sentences is consistent. HT on two sentences implies HT on their disjunction.

Proposition 4.20 (HT on combined sentences)

$$\vdash \ (HT_{ab}(\lambda, \delta) \wedge HT_{ab}(\sigma, \delta)) \ \supset HT_{ab}(\lambda \vee \sigma, \delta) .$$

Proof: First, $HT_{ab}(\lambda, \delta) \wedge HT_{ab}(\sigma, \delta)$ implies $SINC_{ab}(\lambda) \wedge SINC_{ab}(\sigma)$ which is equivalent to $SINC_{ab}(\lambda \wedge \sigma)$ (Proposition 2.3). It implies $SINC_{ab}(\lambda \vee \sigma)$ by ($\mathbf{P}$), ($\mathbf{K_C}$), ($\mathbf{K_B}$), ($\mathbf{N_C}$), ($\mathbf{N_B}$), and ($\mathbf{MP}$). Secondly, $HT_{ab}(\lambda, \delta) \wedge HT_{ab}(\sigma, \delta)$ implies $B_aB_b(\lambda \supset \delta) \wedge B_aB_b(\sigma \supset \delta)$, which implies $B_aB_b((\lambda \vee \sigma) \supset \delta)$ by Proposition 2.1 and ($\mathbf{E_B}$). Hence, the result holds. □

By contrast, $(HT_{ab}(\lambda, \delta) \wedge HT_{ab}(\sigma, \delta)) \supset HT_{ab}(\lambda \wedge \sigma, \delta)$ does not hold in general. This is because a believes that individual sentences λ or σ would lead b to believe δ, but this does not imply that the combined sentence $\lambda \wedge \sigma$ would also lead b to believe δ. So $(HT_{ab}(\lambda, \delta) \wedge (I\text{-})SINC_{ab}(\sigma)) \supset HT_{ab}(\lambda \wedge \sigma, \delta)$ does not hold too.

The following properties hold by the definition of HT.

Proposition 4.21 (HT on contrary sentences)

$$\vdash \ (HT_{ab}(\sigma, \delta) \wedge HT_{ab}(\neg\sigma, \delta)) \ \supset \bot .$$

Proof: $HT_{ab}(\sigma, \delta) \wedge HT_{ab}(\neg\sigma, \delta)$ implies $C_{ab}\sigma \wedge C_{ab}\neg\sigma$, contradiction. □

Proposition 4.22 (introspection)

$$\vdash HT_{ab}(\sigma, \delta) \supset B_a\, HT_{ab}(\sigma, \delta) .$$

Proof: The result holds by Definition 4.6 and the axioms ($\mathbf{4_B}$), ($\mathbf{5_B}$), ($\mathbf{4_{CB}}$), ($\mathbf{5_{CB}}$), and ($\mathbf{4_{IB}}$). □

Proposition 4.23 (HT to oneself)

$$\vdash \ (HT_{aa}(\sigma, \delta) \wedge RB_a(\neg\delta)) \supset \bot .$$

Proof: $HT_{aa}(\sigma,\delta) \wedge RB_a(\neg\delta)$ implies $I_a B_a \delta \wedge \neg I_a \neg B_a \neg\delta$. By $(\mathbf{D_B})$, $(\mathbf{N_I})$, and $(\mathbf{K_I})$, $\vdash \neg I_a \neg B_a \neg\delta \supset \neg I_a B_a \delta$. Hence, $I_a B_a \delta \wedge \neg I_a \neg B_a \neg\delta$ implies $I_a B_a \delta \wedge \neg I_a B_a \delta$ by $(\mathbf{MP})$. Contradiction. □

By contrast, $HT_{aa}(\sigma,\delta) \wedge RB_a(\sigma)$ is consistent because a performs sincere communication on the sentence σ.

One cannot lie (or (I-)BS) and HT on the same sentence. On the other hand, $HT_{ab}(\sigma,\delta)$ implies $WI_{ab}(\neg\delta)$.

Proposition 4.24 (relationship between lie, BS, WI and HT)

1. $\vdash (\, LIE_{ab}(\sigma) \wedge HT_{ab}(\sigma,\delta)\,) \supset \bot$.
2. $\vdash (\, (I\text{-})BS_{ab}(\sigma) \wedge HT_{ab}(\sigma,\delta)\,) \supset \bot$.
3. $\vdash HT_{ab}(\sigma,\delta) \supset WI_{ab}(\neg\delta)$.

Proof: 1. $LIE_{ab}(\sigma) \wedge HT_{ab}(\sigma,\delta)$ implies $B_a\neg\sigma \wedge B_a\sigma$, contradiction. *2.* $(I\text{-})BS_{ab}(\sigma) \wedge HT_{ab}(\sigma,\delta)$ implies $\neg B_a\sigma \wedge B_a\sigma$, contradiction. *3.* $HT_{ab}(\sigma,\delta)$ implies $B_a\neg\delta \wedge \neg C_{ab}\neg\delta \wedge I_a B_b \delta$. By $I_a B_b\delta$, $I_a\neg B_b\neg\delta$ is proved by $(\mathbf{D_B})$, $(\mathbf{N_I})$, $(\mathbf{K_I})$, and $(\mathbf{MP})$. Hence, $HT_{ab}(\sigma,\delta)$ implies $WI_{ab}(\neg\delta)$. □

HT with objectives is defined as follows.

Definition 4.7 (HT with objective) Let a and b be two agents and $\delta, \sigma, \varphi \in \Phi$.

$$O\text{-}HT_{ab}(\sigma,\delta,\varphi) \stackrel{def}{=} I_a B_b\, \varphi \wedge \neg B_a B_b \varphi \wedge B_a B_b(\delta \supset \varphi) \wedge HT_{ab}(\sigma,\delta). \qquad (15)$$

In this case, a provides a *half-truth* to b on σ to reach δ with an *objective* φ.

In (15), a provides a half-truth sentence σ to b with an objective φ if (i) a has an intention to make b believe φ, and (ii) a disbelieves that b believes φ, while (iii) a believes that the believed-false sentence δ leads b to believe φ, then (iv) a provides a half-truth sentence σ to b to make b believe δ. In particular, when $\delta \equiv \varphi$, it holds that $O\text{-}HT_{ab}(\sigma,\varphi,\varphi) \equiv HT_{ab}(\sigma,\varphi)$.

Note that HT already has an objective by itself. A speaker provides a believed-true sentence σ with the intention that a hearer believes a believed-false sentence δ. On the other hand, the sentence φ in Definition 4.7 is not necessarily a believed-false sentence by the speaker. Let us illustrate this by Example 4.1: John informs Mary that he gets a permanent position at a company. John intends Mary to believe that he will get a stable income, which he believes false. John expects that Mary will decide to marry him by this fact. In this case, $O\text{-}HT_{ab}(\texttt{permanent},\texttt{stable},\texttt{marry})$ holds.

5 Maxims for Dishonest Agents

The philosopher Paul Grice introduces the following *maxims for conversation* [35]:

1. (**The maxim of quality**)

- Do not say what you believe to be false.
- Do not say that for which you lack adequate evidence.

2. (**The maxim of quantity**)
   - Make your contribution as informative as is required (for the current purposes of the exchange).
   - Do not make your contribution more informative than is required.
3. (**The maxim of relation**) Be relevant.
4. (**The maxim of manner**) Avoid obscurity of expression; Avoid ambiguity; Be brief (avoid unnecessary prolixity); and Be orderly.

Lies violate the first item of the maxim of quality, and BS violates the second item of it. WI and HT violate the first item of the maxim of quantity. Thus, Grice's maxims of quality/quantity are violated by dishonest agents.[20] In fact, Grice introduces those maxims in his *cooperative principle*, while dishonest speech acts are non-cooperative communication in general. Then our question is whether Grice's maxims can give us any guideline for dishonest agents in their communication. This section provides quantitative and qualitative guidelines for dishonest communication that should ideally be followed by the agents.

5.1 Quantitative maxims

We first consider quantitative guidelines for dishonest communication. Normally one wants to keep his/her dishonesties as small as possible. In lying, for instance, a smaller lie would be considered less sinful than a bigger one from the moral viewpoint. Moreover, from self-interested reasons, a smaller lie would cause less personal discomfort and result in lower criticism or punishment if detected. Kupfer says, "The lie, to his immediate advantage, often results in an overall net loss of freedom in what he can do or say. [...] The liar must be circumspect in his speech and action, guarding against the emergence of his real beliefs. The need to maintain the deception binds him" [38, p.119]. A bigger lie makes the liar less free, which he wants to avoid. From the practical viewpoint, lies make the belief state of a hearer deviate from the objective reality (or, at least from the reality as believed by a speaker) and a bigger lie would increase such deviation. This is undesirable for a speaker because it increases the chance of the lie being detected. Thus, an effective lie is a lie that does not have too much "collateral damage" on a hearer. This intuition is formulated by the following postulate which is meant to have a normative value. Let λ, σ, $\varphi \in \Phi$ such that $\nvdash \lambda \supset \sigma$. Then

$$(\mathbf{P_L}) \qquad B_a(O\text{-}LIE_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}LIE_{ab}(\lambda \wedge \sigma, \varphi) \supset B_b\varphi) \supset \neg O\text{-}LIE_{ab}(\lambda \wedge \sigma, \varphi).$$

$(\mathbf{P_L})$ says that if a believes that two lies λ and $\lambda \wedge \sigma$ are usable for achieving an objective φ, then a does not lie on the bigger one $\lambda \wedge \sigma$. Every (natural or artificial)

[20] Fallis [30] argues that stating a believed-false sentence is considered a lie only if Grice's first maxim of quality is in effect as a norm of conversation.

agent should try to satisfy the postulate $(\mathbf{P_L})$ as well as those discussed further on in this section when communicating with another agent. $(\mathbf{P_L})$ can be satisfied by the agent abstaining from communicating the lie $\lambda \wedge \sigma$ in case he/she believes that a simpler lie λ succeeds in persuading the hearer of believing φ. More specifically, in Example 3.1, if a salesperson a believes that $O\text{-}LIE_{ab}(\texttt{high\_quality}, \texttt{buy})$ and $O\text{-}LIE_{ab}(\texttt{high\_quality} \wedge \texttt{valuable}, \texttt{buy})$ are both effective to persuade a customer to buy a product, then a can satisfy the postulate $(\mathbf{P_L})$ by not lying on the sentence $\texttt{high\_quality} \wedge \texttt{valuable}$.

Similar postulates are considered for BS and WI. Let $\lambda, \sigma, \varphi \in \Phi$ such that $\nvdash \lambda \supset \sigma$. Then

$$(\mathbf{P_{BS}}) \quad B_a(O\text{-}BS_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}BS_{ab}(\lambda \wedge \sigma, \varphi) \supset B_b\varphi) \supset \neg O\text{-}BS_{ab}(\lambda \wedge \sigma, \varphi).$$

$$(\mathbf{P_{WI}}) \quad B_a(O\text{-}WI_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}WI_{ab}(\lambda \wedge \sigma, \varphi) \supset B_b\varphi) \supset \neg O\text{-}WI_{ab}(\lambda \wedge \sigma, \varphi).$$

The postulates $(\mathbf{P_L})$, $(\mathbf{P_{BS}})$ and $(\mathbf{P_{WI}})$ are summarized as the next maxims.

[Quantitative Maxims for Dishonesty]

Maxim I: Lie as little as possible to achieve your objective.

Maxim II: BS as little as possible to achieve your objective.

Maxim III: WI as little as possible to achieve your objective.

These maxims say that it is reasonable (and courteous to a hearer) not to lie, BS, and WI more than absolutely necessary.[21] In HT, on the other hand, this is not necessarily the case. The reason is that providing more information in HT increases the knowledge of a hearer. For a speaker, providing more information implies concealing less information, which alleviates immoral feeling of the speaker. Thus, there seems no reason to prefer a smaller HT that provides less information, so we do not have a maxim mandating it.

5.2 Qualitative maxims

Next we consider qualitative guidelines for dishonest communication. Comparing lies and BS, lies are considered more sinful than BS. This is because a liar intentionally implants wrong beliefs at the hearer, while a bullshitter spits out statements without knowing if they are true. As a result, "people do tend to be more tolerant of bullshit than of lies, perhaps because we are less inclined to take the former as a personal affront" [33, p.50]. This intuition is formulated as follows. Let $\lambda, \sigma, \varphi \in \Phi$ such that $\lambda \not\equiv \sigma$. Then

$$(\mathbf{P_{LB}}) \quad B_a(O\text{-}LIE_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}BS_{ab}(\sigma, \varphi) \supset B_b\varphi) \supset \neg O\text{-}LIE_{ab}(\lambda, \varphi).$$

[21] The maxims I and II are actually in line with the Grice's second quantity maxim "Do not make your contribution more informative than is required".

($\mathbf{P_{LB}}$) says that if a believes that a lie σ and BS λ are both usable for achieving an objective φ, then a does not choose lying on σ.

Next, suppose that an agent believes that both a lie and WI (or both BS and WI) are effective to achieve a goal. In this case, our intuition says that WI is preferable to a lie or BS because WI does not introduce any sentence that is disbelieved by a speaker. The intuition is formulated as follows. Let λ, σ, $\varphi \in \Phi$ such that $\lambda \not\equiv \sigma$. Then

$$(\mathbf{P_{LW}}) \quad B_a(O\text{-}LIE_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}WI_{ab}(\sigma, \varphi) \supset B_b\varphi) \supset \neg O\text{-}LIE_{ab}(\lambda, \varphi).$$

$$(\mathbf{P_{BW}}) \quad B_a(O\text{-}BS_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}WI_{ab}(\sigma, \varphi) \supset B_b\varphi) \supset \neg O\text{-}BS_{ab}(\lambda, \varphi).$$

Finally, HT is considered preferable to lies, BS and WI as an agent communicates a believed-true sentence. Let δ, λ, σ, $\varphi \in \Phi$ such that $\lambda \not\equiv \sigma$. Then we have the following postulates.

$$(\mathbf{P_{LHT}}) \quad B_a(O\text{-}LIE_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}HT_{ab}(\sigma, \delta, \varphi) \supset B_b\varphi) \supset \neg O\text{-}LIE_{ab}(\lambda, \varphi).$$

$$(\mathbf{P_{BHT}}) \quad B_a(O\text{-}BS_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}HT_{ab}(\sigma, \delta, \varphi) \supset B_b\varphi) \supset \neg O\text{-}BS_{ab}(\lambda, \varphi).$$

$$(\mathbf{P_{WHT}}) \quad B_a(O\text{-}WI_{ab}(\lambda, \varphi) \supset B_b\varphi) \wedge B_a(O\text{-}HT_{ab}(\sigma, \delta, \varphi) \supset B_b\varphi) \supset \neg O\text{-}WI_{ab}(\lambda, \varphi).$$

These postulates ($\mathbf{P_{LB}}$), ($\mathbf{P_{LW}}$), ($\mathbf{P_{BW}}$), ($\mathbf{P_{LHT}}$), ($\mathbf{P_{BHT}}$) and ($\mathbf{P_{WHT}}$) are put together as the next maxims.

[Qualitative Maxims for Dishonesty]

Maxim IV: Never lie if you can achieve your objective by BS.[22]

Maxim V: Never lie nor BS if you can achieve your objective by WI.

Maxim VI: Never lie, BS, nor WI if you can achieve your objective by HT.

The qualitative and quantitative maxims provide guidelines that agents should try to satisfy for both moral and self-interested reasons (lower punishments if caught). If we assume that agents try to satisfy the dishonesty maxims, then one can characterize an agent's morality by the worst level of dishonesty he/she is willing to commit in order to achieve a goal. For instance, a lawyer agent might be willing to act on HT (providing only information favorable to his client) or WI (no lawyer would voluntarily provide information against the desirable outcome of his case), but not to BS nor to lie. So if one detects that an agent is performing HT, then one cannot infer that he/she is also willing to WI, BS or lie. However, the other way round, if an agent is willing to lie, then he/she can also be assumed to be willing to BS, WI or to HT. So an agent that is caught on HT (or WI or BS) can perhaps still be trusted not to lie (if trust is the

[22] A similar imperative is mentioned in [33].

Figure 1: Degree of trust

default attitude), but an agent that is caught on lying cannot be trusted at all anymore (Figure 1). In multiagent systems, if agents have implemented the dishonesty maxims, then this can be helpful for reasoning about the possible dishonesty of other agents, and about the extent to which they can still be trusted.

In moral philosophy, dishonest behaviors are often considered morally right when (and only when) they have better consequences than behaving honestly [15].[23] In this sense, "dishonesty" does not necessarily connote "wrongness". The maxims provided in this section just compare dishonest behaviors from communication viewpoints, and do not consider contexts where dishonest behaviors have taken place. For purposes of moral reasoning, this paper mainly focuses on fact finding: has dishonesty taken place? It does not, however, aim to infer whether or not dishonesty may still be morally justifiable, which would be a separate research topic. However, any reasoning about the morality of dishonesty presupposes a theory on what dishonesty is, and what its properties are. It is this theory that we aim to supply in this paper.

6 Discussion

In this section, we first compare different categories of dishonesty formulated so far, and examine their connection to another category of dishonesty, *deception*. We then overview related studies on dishonesty.

6.1 Comparison of different categories of dishonesty

In the previous sections, we have formulated lies, (intentional) BS, withholding information and half-truths, and examined their logical properties. Table 1 compares different categories of dishonesties from the perspective of satisfaction of various conditions discussed so far, and contrasts them with sincere communication. In the table, the properties of **statement**, **addressee**, **truthful** and **intention** stand for the corresponding conditions provided in Section 2.2. The property of **untruthful** stands for the untruthfulness condition provided in Section 3.1. The property of **sincere** means the condition of sincere communication (Definition 2.1). The property of **inability**-$\top$ (resp. **inability**-$\bot$) means the inability to apply valid (resp. contradictory) sentences. The property of **inability**-$\neg$ means the inability on contrary sentences. The property of **inability**-B means the inability on one's own belief. The property of **combination** means that speech acts on individual sentences imply speech on the conjunction of

[23] A famous example of this is "the murderer at the door"—lying to the murderer who asks where his victim has gone [7, 39].

Table 1: Comparison of different categories of dishonesties

	(I-)SINC	**LIE**$^{(B)}$	**(I-)BS**	**WI**	**HT**
statement	$\surd$	$\surd$	$\surd$		$\surd$
addressee	$\surd$	$\surd$	$\surd$		$\surd$
truthful	$\surd$				$\surd$
untruthful		$\surd$			
intention	$(\surd)^*$	$\surd^\dagger$	$(\surd)^\ddagger$	$\surd$	$\surd$
sincere	$\surd$				$\surd$
inability-$\top$		$\surd$ (3.3)	$\surd$ (4.3)	$\surd$ (4.10)	
inability-$\bot$	$\surd$ (2.2)	$\surd$ (3.3)	$\surd$ (4.3)	$\surd$ (4.10)	$\surd$ (4.19)
inability-$\neg$	$\surd$ (2.4)	$\surd$ (3.6)	$\surd$ (4.4)	$\surd$ (4.13)	$\surd$ (4.21)
inability-B			$\surd$ (4.1)		
combination	$\surd$ (2.3)	$\surd$ (3.4)		$\surd$ (4.11)	
introspection	$\surd$ (2.7)	$\surd$ (3.9)	$\surd$ (4.6)	$\surd$ (4.14)	$\surd$ (4.22)
self-contradiction		$\surd$ (3.10)		$\surd$ (4.15)	$\surd$ (4.23)

$*$: The result holds for **I-SINC**. †: The result holds for **LIE**. ‡: The result holds for **I-BS**.

those sentences. The property of **introspection** means that one has a belief of his/her act. The property of **self-contradiction** means that a speech act to oneself leads to contradiction in the presence of the rationally-balanced condition (Definition 2.3). The mark $\surd$ means satisfaction of each condition, and the attached number indicates the corresponding proposition. From the table, we can observe the following facts on different categories of dishonesty. **(i)** The four dishonesties share the properties of inability-$\bot$, inability-$\neg$, and introspection. **(ii)** Lies are untruthful, while others are not. **(iii)** Lies and WI satisfy the combination property, while others do not.[24] **(iv)** BS is the only category that has both intentional and unintentional cases, and is also the only category that does not satisfy self-contradiction. Moreover, BS is the only case that cannot act on one's belief. **(v)** WI is the only category that requires neither statement nor addressee. **(vi)** HT is the only category that does not satisfy inability-$\top$, and is also the only case that is truthful and sincere. So each category can be distinguished by the above properties from other categories. The comparison also explains the intuition behind the qualitative maxims in Section 5.2. From the viewpoint of untruthfulness, LIE$^{(B)}$ is worse than all other dishonesties. From the viewpoint of statement, LIE$^{(B)}$ and (I-)BS are worse than WI and HT because LIE$^{(B)}$ and (I-)BS make statements that is not truthful. From the viewpoint of sincerity, LIE$^{(B)}$, (I-)BS, and WI are worse than HT.

6.2 Deception

Deception is an act whereby one person causes another person to have a false belief. In [41], Mahon says that "Philosophers agree that 'deceive' is a success or an achievement

[24] Although HT cannot be combined on conjunction, it can be combined on disjunction (Proposition 4.20).

verb, and that an act of deceiving is a perlocutionary act. Whether or not an act of deceiving has occurred depends on whether or not a particular effect—normally, the having of a false belief—has been produced in another; if no such effect has been produced in another, then no deceiving has occurred." In this respect, deception is different from any category of dishonesties studied in this paper. According to their definitions, whether or not a speaker lies (BS, WI, or HT) depends only on the belief and intention of a speaker, and is independent of the effect of the action.

Since deception involves a success or an achievement of the act, its formulation requires a logic that can express causation and effects as well as belief and intention. Sakama and Caminada [56] formalize different forms of deception using a modal logic of action and belief developed by Pörn [52]. They provide logical formulation of *deception by commission* and *deception by omission*, that were originally described by Chisholm and Feehan [19]. They investigate formal properties of eight different categories of deception and discuss their relationship to lying. Here we informally characterize deception based on dishonesty studied in this paper by introducing a *causal relation* to the logic. A sentence "$p \Rightarrow q$" is read as "q is a consequence of p". It satisfies the axiom $(p \Rightarrow q) \supset (p \wedge q)$ [61]. Semantically, $p \Rightarrow q$ is true at a possible world if and only if q is true in every world selected in terms of the given world in which p is true.[25] With this extension, we could define different types of deception based on lies, I-BS, WI and HT as follows.

$$DecLie_{ab}(\sigma) \stackrel{def}{=} LIE_{ab}(\sigma) \Rightarrow B_b\sigma\,.$$
$$DecBS_{ab}(\sigma) \stackrel{def}{=} I\text{-}BS_{ab}(\sigma) \Rightarrow B_b\sigma\,.$$
$$DecWI_{ab}(\sigma) \stackrel{def}{=} WI_{ab}(\sigma) \Rightarrow B_b\neg\sigma\,.$$
$$DecHT_{ab}(\sigma,\delta) \stackrel{def}{=} HT_{ab}(\sigma,\delta) \Rightarrow B_b\delta\,.$$

$DecLie$ is *deception by lying*, $DecBS$ is *deception by bullshitting*, $DecWI$ is *deception by withholding information*, and $DecHT$ is *deception by half-truths*. The intuitive reading of $DecLie$ is that an agent a deceives another agent b on a sentence σ if a lies to b on σ which brings about b's believing σ. $DecBS$, $DecWI$, and $DecHT$ have similar meanings. Note that the definitions do not describe the belief state of a hearer b before deceptive actions have taken place. If b originally disbelieves σ, then $LIE_{ab}(\sigma)$ contributes causally to b's acquiring the belief in σ. Else if b originally believes σ, then $LIE_{ab}(\sigma)$ contributes causally to b's continuing in the belief in σ. Although we do not pursue the details of the logic and formal properties of deception in this paper, we would still like to emphasize the difference between deception and the other categories of dishonesty. By definition, deception succeeds if a dishonest act of an agent makes the belief state of a hearer conform with the intention of a speaker. On the other hand, *attempted deception* does not always succeed. For instance, attempted deception by lying fails if $LIE_{ab}(\sigma) \not\Rightarrow B_b\sigma$ holds (i.e., a lies to b on σ while it does *not* bring about b's believing σ). As such, there are two cases of lying: the one is lying that attempts but fails to deceive, and the other is lying that deceives.[26] Similar distinctions are made

[25] We refer to [56, 61] for the formal properties of the causality operator $\Rightarrow$.

[26] Carson [15] also considers the third case: lying without attempted deception. This is because he does

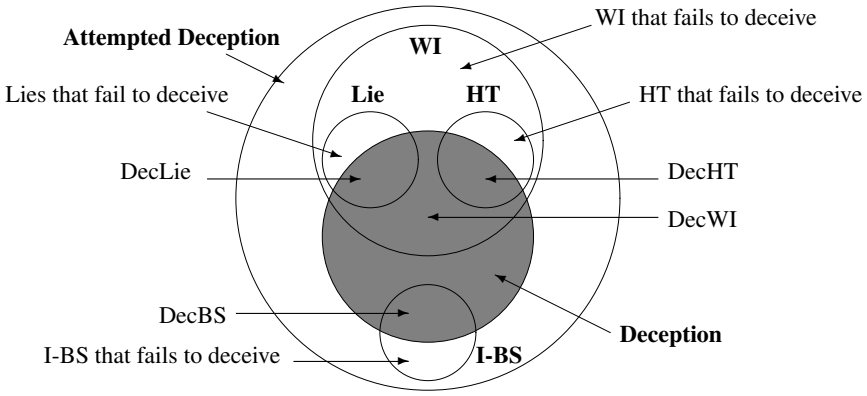

Figure 2: Lie, I-BS, WI, HT, deception and attempted deception

for I-BS, WI, and HT. Thus, deception is the successful case of attempted deception [15]. The relationship between lies, I-BS, WI, HT, deception and attempted deception is illustrated in Figure 2. Note that lies imply WI (Proposition 4.16) and HT implies WI (Proposition 4.24). On the other hand, lies and (I-)BS, lies and HT, (I-)BS and WI, or (I-)BS and HT, are all mutually exclusive (Propositions 4.7, 4.17 and 4.24).

6.3 Related work

Some attempts have been made to formulate dishonesty using modal logic.

O'Neill [48] provides logical definitions of lies and deception based on the logic of intentional communication [23]. He uses four different modalities of belief, intention, common belief, and communication. A lie is defined as $Lie_{ab}\,\sigma = C_{ab}\,\sigma \wedge B_a \neg\sigma$ where C_{ab} means that "a person a communicates a proposition σ to a person b". Here C_{ab} is called *assertive communication* satisfying the relation $C_{ab}\,\sigma \equiv B^*\,I_a\,(C_{ab}\,\sigma \wedge B_b B_a\,\sigma)$ where $B^*\varphi$ means that φ is a common belief between a and b. He also defines the so-called "talking through your hat" as $Hat_{ab}\,\sigma = C_{ab}\,\sigma \wedge \neg B_a\sigma$ and deception as $Dec_{ab}\,\sigma = I_a B_b\sigma \wedge B_a\neg\sigma \wedge B_b\sigma$. $Lie_{ab}\,\sigma$ represents that a speaker a intends to get a hearer b to believe not necessarily σ but only that a believes σ. This corresponds to the intention to deceive about truthfulness of $LIE^B_{ab}(\sigma)$, and is different from the intention to deceive about truth of $LIE_{ab}(\sigma)$. $Hat_{ab}\,\sigma$ is close to $BS_{ab}(\sigma)$ but has the same intentional nature as $Lie_{ab}\,\sigma$. $Dec_{ab}\,\sigma$ represents that deception happens when a intends to make b believe a believed-false sentence σ and b believes it. However, the definition does not represent that b comes to have a false belief σ as a result of some action of a. According to $Dec_{ab}\,\sigma$, a deceives b when b believes σ regardless of any action of a, which is rather odd. The primary interest of [48] is to formulate various types of speech acts in an epistemic logic, and he does not provide comparative analyses of different categories of dishonesties.

not consider that intention to deceive is necessary for lying. In our definition of lying, however, a speaker intends to deceive a hearer, so we do not consider the third case here.

Firozabadi *et al.* [31] formulate *fraud* using modal operators for obligation, action and belief. According to their definition, fraud is a situation where an agent violates one of its obligations and also deceives another agent that the obligation is fulfilled. An action of an agent is considered deceptive if he/she either does not have a belief about the truth value of some proposition but makes another agent believe that the proposition is true or false, or he/she believes that the proposition is true/false but makes another agent believe the opposite. These two cases are formally defined in [31] as: $\neg B_a\varphi \wedge E_aB_b\varphi$ and $B_a\neg\varphi \wedge E_aB_b\varphi$ where $E_a\psi$ means "a brings it about that ψ". These definitions do not mention any means that an agent uses to bring about a result. The authors also consider that an agent who does not succeed in his/her attempt to deceive another agent is still a deceptive agent. Such cases, which we call attempted deception and distinguish them from deception, are formally represented by $\neg B_a\varphi \wedge H_aB_b\varphi$ and $B_a\neg\varphi \wedge H_aB_b\varphi$, where $H_a\psi$ means that "an agent a attempts to bring about ψ, not necessarily successful". Their goal is to characterize specific types of fraud situations that may occur in organized interactions like trade procedures. Firozabadi and Jones [32] define lying in terms of basically the same action logic as [31] and formulate trust of an agent. Suppose that an agent a, by saying something or by sending a particular document, gets another agent b to believe that σ. Let Δ denote "the document or information is delivered to an agent b". Then they define lying as $\neg B_a\,\sigma \wedge E_a(\Delta, B_b\sigma)$ which represents that a disbelieves σ, and that a makes b believe that σ by bringing it about that Δ. The definition satisfies none of the four necessary conditions of lying in Section 3.1, while it requires the success of lying. As a result, it is not considered as the standard definition of lying.

Pan *et al.* [49] provide an axiomatic system for lies using a multi-modal logic. According to their definition, a speech act x is a lie if a speaker intends a hearer by performing x to form some beliefs that the speaker regards as false. Then they consider two different types of lies based on whether or not a speech act x satisfies the *sincerity principle*. Let $S(x)$ stand for "utterance x is sincere". A speech act x is a *direct lie* if a speaker does not believe $S(x)$ and intends a hearer by performing x to believe $S(x)$. A speech act x is an *indirect lie* if a speaker does believe $S(x)$ but does not believe some proposition φ and intends the hearer to believe φ by performing x. In direct lies, a speaker does not intend to make a hearer believe a particular sentence, but intends to make a hearer believe that the speaker is sincere. On the other hand, in indirect lies, a speaker intends to make a hearer believe a particular sentence under the assumption that the speaker believes that he/she conforms to the rules of sincerity principle. They formulate these two types of lying using their logic, but they just provide definitions and do not investigate formal properties.

Baltag and Smets [6] introduce a logic of conditional doxastic actions. According to their formulation, the action of public successful lying is characterized by an action plausibility model involving two actions $Lie_a(\sigma)$ and $True_a(\sigma)$. The former represents an action in which an agent a publicly lies that she knows σ while in fact she does not know it. The latter represents an action in which a makes a public truthful announcement that she knows σ. They have preconditions $\neg K_a\sigma$ and $K_a\sigma$, respectively. If a hearer b trusts the agent a and he is inclined to believe the lie, then the situation is represented by the pre-order $True_a(\sigma) <_b Lie_a(\sigma)$ which means that it is more plausible to b that a is telling the truth rather than lying. If a hearer already knows that σ is

false, however, the action $Lie_a(\sigma)$ does not succeed. Such a condition is formulated as an action's *contextual appearance*. Note that the precondition $\neg K_a\sigma$ of $Lie_a(\sigma)$ represents the ignorance of σ and is different from the untruthfulness condition. In their formulation, lying is given as a basic epistemic action in an action plausibility model, so it has no logical definition. Their goal is formulating multiagent belief updates with actions by combining belief revision and dynamic epistemic logic.

Tzouvaras [73] formulates different types of lies and studies the problem of the "Liar Paradox". He introduces a propositional multi-modal logic KU having two modalities: $K\varphi$ ("φ is known") and $U\varphi$ ("φ is uttered"), while it has no modality representing intention. The axiomatic system for K is S5 and the one for U is KD45. He then introduces three definitions of lying:

$$L_1\psi := (\neg K\psi \wedge U\psi) \vee (\neg K\neg\psi \wedge U\neg\psi)$$
$$L_2\psi := (K\psi \wedge U\neg\psi) \vee (K\neg\psi \wedge U\psi)$$
$$L_3\psi := (U\psi \vee U\neg\psi) \wedge (\neg K\psi \wedge \neg K\neg\psi).$$

In L_1, a speaker utters a fact that is not known to be true. In L_2, a speaker utters a false fact. In L_3, a speaker utters a fact that is known to be neither true nor false. By definition, L_3 implies L_1. L_1 and L_3 do not satisfy the untruthfulness condition. On the other hand, L_2 requires the statement to be false because $K\psi$ implies the truth of ψ and $K\neg\psi$ implies the falsity of ψ. As argued in Section 3.1, however, this is considered too strong, since the actual falsity of a sentence is not necessarily required in lying. Moreover, none of them satisfies the intention condition. Using the relation $K_a\psi \supset B_a\psi$ between knowledge and belief [37], our definition of bullshit implies L_3. Tzouvaras analyzes the Liar Paradox using these definitions. Let σ be the sentence that "I am lying". Then the sentence $L\sigma$ is called *intentional lying* where L is one of L_1, L_2 or L_3. In this case, the liar is in *intentional mode* iff $\sigma \equiv L\sigma$ in that particular situation. Tzouvaras shows that the paradox is resolved when a speaker is in intentional mode, for instance, $\sigma \equiv L_1\sigma$ and $U\sigma$ just imply $\sigma \wedge \neg K\sigma$.

In his formulation of trustful agents, Liau [40] defines that an agent i is an *intentional liar* if $U_i\psi \wedge B_i\neg\psi$, while i is an *irresponsible liar* if $U_i\psi \wedge \neg B_i\psi$. Here, $U_i\psi$ means that an agent i utters ψ. Then an agent i is *honest* if he/she is not an irresponsible liar (i.e., $U_i\psi \supset B_i\psi$ for all ψ). Note that if an agent i is honest, he/she is not an intentional liar (i.e., $(U_i\psi \supset B_i\psi) \supset (U_i\psi \supset \neg B_i\neg\psi)$). He argues that an ideal form of trust, which he calls *cautious trust*, should satisfy two conditions: first, an agent i believes that if another agent j tells him ψ then j himself believes ψ (*honesty*); and second, i believes that if j believes ψ then ψ in fact holds (*capability*). Such cautious trust is formally defined as $T^c_{ij}\psi \stackrel{def}{=} B_i((U_j\psi \supset B_j\psi) \wedge (B_j\psi \supset \psi))$. He addresses that this is a very strict requirement for i to trust j on ψ, then he uses a weaker notion of trust T_{ij} as a primitive operator in his logic.

Van Ditmarsch *et al.* [75] formulate a logic of lying in public discourse. They use a propositional dynamic logic extended with manipulative updates, lies and announcements. Different from ours, the act of lying is provided as a modality in their logic. In notation, $[¡\varphi]\sigma$ means that a formula σ is true after lying public announcement of φ, which is contrasted with $[!\varphi]\sigma$ representing a truthful public announcement. Van

Ditmarsch [76] also studies dynamic aspects of lying and bluffing using dynamic epistemic logic. The primary interest of [75, 76] is not in precisely formulating what is lying, but in modelling how the belief of an agent is changed by (un)truthful announcements. The study [75] also provides a game-theoretical analysis of lying as an optimal strategy in a two-person game. There have been game theoretic approaches to capture the phenomenon of lying and deception (e.g. [24, 28]). The goal of these studies is analyzing the effect of lying and deception in strategic games, rather than understanding what is lying and deception. Sakama *et al.* [55] provide logical formulation of lies, BS and deception. They introduce different types of lies, called *deductive lies* and *abductive lies*, and investigate formal properties. Deception in [55] is similar to half-truth of this paper, while these two notions are distinguished in this paper. In this paper we consider two definitions of lies that are considered to be the best ones by [42], while there is a number of different definitions of lies in the literature of philosophy. Sakama [57] provides a formal analysis of twelve definitions of lies that have been proposed in the philosophical literature and were analyzed in an informal way by Mahon [42]. Isaac and Bridewell [36] consider a situation that a hearer believes a proposition p and a speaker utters the negation of p. In this case, the hearer infers the speaker's *ulterior motive* behind the utterance for deciding further actions to take. They consider four different categories of dishonesty: lying, bullshit, paltering and pandering. *Paltering* involves speaking truthfully with the intent to deceive, which is similar to HT in this paper. *Pandering* resembles bullshit and the speaker does not care about the truth value of his utterance (although he may possibly *know* it). In their approach, the hearer reads the speaker's mind from a dialogue content and classifies possible states of dishonest acts by the speaker. We do not study ulterior motives behind dishonesty and mindreading from the hearer's viewpoint in this paper. They introduce a framework for identifying deceptive entities (FIDE) for a mindreading system to detect deception, which is not based on a formal logic, however. Clark [20] develops a *lying machine* which uses *mental models theory* to exploit cognitive illusions, highly fallible heuristics and biases in human reasoning. The machine incorporates those illusions into its arguments, and articulates the illusory arguments with the intent to deceive its audience. The study provides empirical evidence that the machine reliably deceives ordinary humans.

Some studies show potential utility of dishonest reasoning in applications. Bonatti *et al.* [8] study databases that may provide users with incorrect answers to preserve security in a multi-user environment. They introduce a propositional modal logic to reason about databases, secrets, and users' beliefs. Sklar *et al.* [66] formulate lying in an argument-based dialogue game. They describe an education system that presents students with a problem and a false solution, and when they object to the solution, asking them to justify their reaction. De Rosis *et al.* [27] study how agents can deceive within a probabilistic framework for representing mental states. Their theory is applied to a simplified version of Turing's imitation game and different types of deceptive strategies are implemented. Zlotkin *et al.* [81] study negotiation in which agents may lie. The study shows how an agent can benefit himself/herself by effectively lying in a deal of cooperative planning. Castelfranchi *et al.* [17] present the utility of lies and deception in order to obtain help or delegate tasks in cooperative activities. Wagner and Arkin [78] develop a robot that uses a deceptive communication to escape from an enemy robot. A robot determines whether or not a situation warrants deception based

on an expected outcome in a game theory. Son *et al.* [69] provide a logical framework for negotiation among dishonest agents. They show how intentionally false or inaccurate information can be effectively used in the process of negotiation to have desired outcomes by agents. Caminada [11] provides a comparative study between lies, bullshit and deception, and discusses how an intelligent agent could behave dishonestly to win a debate in formal argumentation systems. Rahwan *et al.* [53] characterize the strategy-proofness (i.e., truth-telling being a dominant strategy equilibrium) in formal argumentation systems when agents may hide and/or lie about arguments. Sakama [59] provides a formal model of debate games in which a player may provide false or incorrect arguments as a tactic to win the game. Staab and Caminada [70] design and implement an MAS-based software simulator and observe how the incentives for dishonesty emerge in self-interested agents aiming to optimize their economic performance. Caminada *et al.* [12] show that lying can promote the social welfare as well as increase personal utilities of an agent in situations of argumentation-based judgment aggregation.

7 Conclusion

Lying and other categories of dishonesty have been studied in philosophy and elsewhere, while a logical foundation of dishonesty is a topic that has received relatively little attention. We provided logical accounts of various categories of dishonesty and analyzed their formal properties. The abstract framework proposed here will need to be extended in subsequent work, but is as it stands capable of capturing the declarative kernels of many forms of human dishonesty. Our research does not aim to provide new philosophical insights on dishonesty, but to turn conceptually defined notions in philosophy into a framework based on formal logic. Although some formal properties were provided, the strength of the current paper is aimed to be conceptual rather than purely technical. The maxims for dishonest agents can be seen as having a normative value, and should ideally be implemented for individual agents in multiagent systems. Dishonest behavior involves complex reasoning about a relationship between a speaker and a hearer, their belief states, and context in which a speech act is performed. The current study serves as a kind of base-level and would contribute to opening the topic. An important limitation of the current work is that it is built on top of the Kripke-style possible-world semantics, and therefore inherits some properties that are not very realistic (for instance, logical omniscience) when providing a formal semantics for concepts like beliefs, intentions and communication.[27] Formalizing dishonesty in a more expressive logic than we have currently implemented is a research challenge for the future. We focus on declarative aspects of dishonest reasoning in this paper, while its computational aspects are considered in [58] that introduces *logic programs with disinformation* to represent and reason with dishonesty.

Understanding when an agent behaves dishonestly and how dishonest reasoning is performed by agents is useful in identifying systems that mislead users, and providing ways to protect users from being deceived. A well-designed protocol that can

[27] "Indeed, knowledge, belief, desire, intention, provability, etc., all receive the exact same formal analysis in possible-world semantics" [4].

detect deceptive agents and distinguish disinformation is needed in an open distributed environment like the Internet where a lot of intentionally wrong and misleading information exists [44]. The issue of detecting dishonest agents and preventing a success of deception is not handled in this paper but is an important research issue. Another topic of interest is modelling the emergence of dishonesty in social environments. Some studies show that dishonesty and credulity are behaviors that evolved by natural selection [54, 68]. A recent study shows that artificial robots which compete for food learn to conceal food information [46]. The study [60] models children's acquisition of dishonesty using machine learning techniques, while there is still much to be done.

Acknowledgments We thank the reviewers for their valuable comments on the manuscript. The second author has been supported by the National Research Fund, Luxembourg (LAAMI project) and by the Engineering and Physical Sciences Research Council (EPSRC, UK), grant ref. EP/J012084/1 (SAsSy project).

References

[1] J. E. Adler. Lying, deceiving, or falsely implicating. *Journal of Philosophy*, **94**, 435–452, 1997.

[2] M. Anderson and S. L. Anderson (eds.). *Machine Ethics*. Cambridge University Press, 2011.

[3] D. Ariely. *The (honest) truth about dishonesty: how we lie to everyone—especially ourselves*, HarperCollins Publishers, 2012.

[4] K. Arkoudas and S. Bringsjord. Propositional attitudes and causation. *Journal of Software and Informatics*, **3**, 47–65, 2009.

[5] Augustine. Lying. In: *Treatises on Various Subjects, Fathers of the Church*, vol. 56, pp. 45–110, 1952.

[6] A. Baltag and S. Smets. The logic of conditional doxastic actions: a theory of dynamic multi-agent belief revision. In: *Proceedings of the ESSLLI Workshop on Rationality and Knowledge*, 2006.

[7] S. Bok. *Lying: Moral Choice in Public and Private Life*. Vintage, 1999.

[8] P. A. Bonatti, S. Kraus and V. S. Subrahmanian. Foundations of secure deductive databases. *IEEE Transactions on Knowledge and Data Engineering*, **7**, 406–422, 1995.

[9] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, 1987.

[10] P. Bronson. Learn to lie. *New York Magazine*, February, 2008.

[11] M. Caminada. Truth, lies and bullshit, distinguishing classes of dishonesty. In: *Proceedings of the IJCAI Workshop on Social Simulation*, 2009.

[12] M. Caminada, G. Pigozzi and M. Podlaszewski. Manipulation in group argument evaluation. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 121–126, 2011.

[13] R. E. Carrion, J. P. Keenan and N. Sebanz. A truth that's told with bad intent: an ERP study of deception. *Cognition*, **114**, 105–110, 2010.

[14] T. L. Carson. The definition of lying. *Noûs*, **40**, 284–306, 2006.

[15] T. L. Carson. *Lying and deception: theory and practice*. Oxford University Press, 2010.

[16] C. Castelfranchi. Artificial liars: why computers will (necessarily) deceive us and each other? *Ethics and Information Technology*, **2**, 113–119, 2000.

[17] C. Castelfranchi, R. Falcone and F. De Rosis. Deceiving in Golem: how to strategically pilfer help. In: *Trust and deception in virtual societies*, pp. 91–109, Kluwer Academic Publishers, 2001.

[18] B. F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, Cambridge, 1980.

[19] R. M. Chisholm and T. D. Feehan. The intent to deceive. *Journal of Philosophy*, **74**, 143–159, 1977.

[20] M. Clark. Cognitive illusions and the lying machine: a blueprint for sophistic mendacity. Doctoral Dissertation, Rensselaer Polytechnic Institute, Troy, NY, USA, 2010. A short article: "Mendacity and deception: uses and abuses of common ground", *Proc. AAAI Fall Symposium,* 2011.

[21] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, **42**, 213–261, 1990.

[22] G. A. Cohen. Deeper into bullshit. In: S. Buss and L. Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt*, pp. 321–339, MIT Press, 2002.

[23] M. Colombetti. A modal logic of intentional communication. *Mathematical Social Sciences*, **38**, 171–196, 1999.

[24] V. P. Crawford. Lying for strategic advantage: rational and boundedly: rational misrepresentation of intentions. *American Economic Review*, **93**, 133–149, 2003.

[25] R. Demos. Lying to oneself. *Journal of Philosophy*, **57**, 588–595, 1960.

[26] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer and J. A. Epstein. Lying in everyday life. *Journal of Personality and Social Psychology*, **70**, 979–995, 1996.

[27] F. de Rosis, V. Carofiglio and G. Grassano. Can computers deliberately deceive? A simulation tool and its application to Turing's imitation game. *Computational Intelligence*, **19**, 235–263, 2003.

[28] D. Ettinger and P. Jehiel. A theory of deception. *American Economic Journal: Macroeconomics*, **2**, 1–20, 2010.

[29] R. Fagin, J. Y. Halpern, Y. Moses and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[30] D. Fallis. What is lying? *Journal of Philosophy*, **106**, 29–56, 2009.

[31] B. S. Firozabadi, Y. H. Tan and R. M. Lee. Formal definitions of fraud. In: P. McNamara and H. Prakken (eds). *Norms, Logics and Information Systems – New Studies in Deontic Logic and Computer Science*, pp. 275–288, IOS Press, 1999.

[32] B. S. Firozabadi and A. J. I. Jones. On the characterisation of a trusting agent – aspects of a formal approach. In: C. Castelfranchi and Y. H. Tan (eds.), *Trust and Deception in Virtual Societies*, Kluwer Academic Publishers, pp. 157–168, 2001.

[33] H. G. Frankfurt. *On Bullshit*. Princeton University Press, 2005.

[34] J.-G. Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and Information Technology*, **9**, 39–47, 2006.

[35] H. P. Grice. *Studies in the Ways of Words*. Harvard University Press, 1989.

[36] A. M. C. Isaac and W. Bridewell. Mindreading deception in dialog, *Cognitive Systems Research*, **28**, 12–19, 2014.

[37] S. Kraus and D. Lehmann. Knowledge, belief and time. *Theoretical Computer Science*, **58**, 155–174, 1988.

[38] J. Kupfer. The moral presumption against lying. *Review of Metaphysics*, **36**, 103–126, 1982.

[39] I. Leslie. *Born Liars—why we can't live without deceit*. Quercus, London, 2011.

[40] C-J. Liau. Belief, information acquisition, and trust in multi-agent systems – a modal logic formulation. *Artificial Intelligence*, **149**, 31–60, 2003.

[41] J. E. Mahon. A definition of deceiving. *Journal of Applied Philosophy*, **21**, 181–194, 2007.

[42] J. E. Mahon. Two definitions of lying. *Journal of Applied Philosophy*, **22**, 211–230, 2008.

[43] J. E. Mahon. The definition of lying and deception. *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/lying-definition/ , 2008.

[44] A. P. Mintz. *Web of Deception: Misinformation on the Internet*. Information Today, 2002.

[45] R. W. Mitchell and N. S. Thompson (eds.). *Deception: perspectives on human and nonhuman deceit*. SUNY Press, 1986.

[46] S. Mitri, D. Floreano and L. Keller. The evolution of information suppression in communicating robots with conflicting interests. In: *Proceedings of National Academy of Sciences 106(37)*, pp. 15786–15790, 2009.

[47] G. E. Moore. Moore's paradox. In T. Baldwin (ed.), *G. E. Moore: Selected Writings*, pp. 207–212, Routledge, 1993.

[48] B. O'Neill. A formal system for understanding lies and deceit. *Jerusalem Conference on Biblical Economics*, 2003.

[49] Y. Pan, C. Cao and Y. Sui. A formal system for lies based on speech acts in multi-agent systems. In: *Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pp. 228–234, 2007.

[50] L. M. Pereira and A. Saptawijaya. Modelling morality with prospective logic. In M. Anderson and S. L. Anderson, (eds.), *Machine Ethics*, pp. 398–421, Cambridge University Press, 2011.

[51] J. Pitt and E. H. Mamdani. Some legal aspects of inter-agent communication: from the sincerity condition to 'ethical' agents. In: *Proceedings of Issues in Agent Communication. Lecture Notes in Artificial Intelligence* 1916, pp. 46–62, Springer-Verlag, 2000.

[52] I. Pörn. On the nature of social order. In J. E. Fenstad et al. (eds.), *Logic, Methodology, and Philosophy of Science*, VIII, Elsevier, 1989.

[53] I. Rahwan, K. Larson and O. Tohmé. A characterisation of strategy-proofness for grounded argumentation semantics. In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 251–256, 2009.

[54] J. T. Rowell, S. P. Ellner and H. K. Reeve. Why animals lie: how dishonesty and belief can coexist in a signaling system. *American Naturalist*, **168**(6), E180–E204, 2006.

[55] C. Sakama, M. Caminada and A. Herzig. A logical account of lying. In: *Proceedings of the 12th European Conference on Logics in Artificial Intelligence (JELIA), Lecture Notes in Artificial Intelligence* 6341, pp. 286–299, Springer-Verlag, 2010.

[56] C. Sakama and M. Caminada. The many faces of deception. In: *Proceedings of the Thirty Years of Nonmonotonic Reasoning (NonMon@30)*, Lexington, KY, USA, October 2010.

[57] C. Sakama. Logical definitions of lying. In: *Proceedings of the 14th International Workshop on Trust in Agent Societies (TRUST)*, Taipei, Taiwan, May 2011.

[58] C. Sakama. Dishonest reasoning by abduction. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1063–1068, 2011.

[59] C. Sakama. Dishonest arguments in debate games. In: *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA), Frontiers in Artificial Intelligence and Applications* 245, IOS Press, pp. 177–184, 2012.

[60] C. Sakama. Learning dishonesty. In: *Proceedings of the 22nd International Conference on Inductive Logic Programming (ILP), Lecture Notes in Artificial Intelligence* 7842, pp. 225–240, Springer-Verlag, 2012.

[61] G. Sandu. Formal Logic of Action. Licentiate Thesis, University of Helsinki, 1986.

[62] J. R. Searle. *Speech Acts*. Cambridge University Press, 1969.

[63] J. R. Searle. *Expression and meaning. Studies in the theory of speech acts*. Cambridge University Press, 1979.

[64] F. A. Siegler. Lying. *American Philosophical Quarterly* **3**(2), 128–136, 1966.

[65] M. P. Singh. Agent communication languages: rethinking the principles. *IEEE Computer*, December, pp. 40–47, 1998.

[66] E. Sklar, S. Parsons and M. Davies. When is it okay to lie? A simple model of contradiction in agent-based dialogues. In: *Argumentation in Multi-Agent Systems (ArgMAS), Lecture Notes in Computer Sciences* 3366, pp. 251–261, Springer-Verlag, 2005.

[67] D. L. Smith. *Why we lie: the evolutionary roots of deception and the unconscious mind*. St. Martin's Griffin, 2007.

[68] E. Sober. The primacy of truth-telling and the evolution of lying. In: *From a Biological Point of View Essays in Evolutionary Philosophy*, Cambridge Studies in Philosophy and Biology, pp. 71–92, 1994.

[69] T. C. Son, E. Pontelli, N.-H. Nguyen and C. Sakama. Formalizing negotiations using logic programming. *ACM Transactions on Computational Logic*, **15**(2), Article No.12, 2014.

[70] E. Staab and M. Caminada. On the profitability of incompetence. In: *Multi-Agent-Based Simulation XI, Lecture Notes in Computer Science* 6532, pp. 76–92, Springer-Verlag, 2011.

[71] R. Trivers. *The Folly of Fools: the logic of deceit and self-deception in human life*. Basic Books, 2011.

[72] A. M. Turing. Computing machinery and intelligence. *Mind* 59:433–460, 1950.

[73] A. Tzouvaras. Logic of knowledge and utterance and the liar. *Journal of Philosophical Logic*, **27**, 85–108, 1998.

[74] W. van der Hoek. Systems for knowledge and belief. *Journal of Logic and Computation*, **3**, 173–195, 1993.

[75] H. van Ditmarsch, J. van Eijck, F. Sietsma and Y. Wang. On the logic of lying. *Games, Actions and Social Software, Texts in Logic and Games*, LNAI-FoLLI, Springer-Verlag, 2011.

[76] H. van Ditmarsch. Dynamics of lying. *Synthese*, Springer-Verlag, 2013.

[77] J. M. Vincent and C. Castelfranchi. On the art of deception: how to lie while saying the truth. In: *Proceedings of the Conference on Pragmatics*, Studies in Language Companion Series 7, pp. 749–777, 1979.

[78] A. R. Wagner and R. C. Arkin. Acting deceptively: providing robots with the capacity for deception. *Journal of Social Robotics*, **3**, 5–26, 2011.

[79] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

[80] K. Warwick and H. Shah. Effects of lying in practical Turing tests. *AI and Society*, 2014.

[81] G. Zlotkin and J. S. Rosenschein. Incomplete information and deception in multi-agent negotiation. In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 225–231, 1991.